

# Context-aware learning for finite mixture models

Serafeim Perdikis

Robert Leeb

Ricardo Chavarriaga

José del R. Millán

*Chair in Brain-Machine Interface*

*Center for Neuroprosthetics*

*Institute of Bioengineering*

*School of Engineering*

*École Polytechnique Fédérale de Lausanne*

*Campus Biotech H4, Chemin des Mines 9*

*CH-1202, Geneva, Switzerland*

SERAFEIM.PERDIKIS@EPFL.CH

ROBERT.LEEB@EPFL.CH

RICARDO.CHAVARRIAGA@EPFL.CH

JOSE.MILLAN@EPFL.CH

**Editor:** None

## Abstract

This work introduces algorithms able to exploit contextual information in order to improve maximum-likelihood (ML) parameter estimation in finite mixture models (FMM), demonstrating their benefits and properties in several scenarios. The proposed algorithms are derived in a probabilistic framework with regard to situations where the regular FMM graphs can be extended with context-related variables, respecting the standard expectation-maximization (EM) methodology and, thus, rendering explicit supervision completely redundant. We show that, by direct application of the missing information principle, the compared algorithms' learning behaviour operates between the extremities of supervised and unsupervised learning, proportionally to the information content of contextual assistance. Our simulation results demonstrate the superiority of context-aware FMM training as compared to conventional unsupervised training in terms of estimation precision, standard errors, convergence rates and classification accuracy or regression fitness in various scenarios, while also highlighting important differences among the outlined situations. Finally, the improved classification outcome of contextually enhanced FMMs is showcased in a brain-computer interface application scenario.

**Keywords:** context-awareness, semi-supervised learning, probabilistic labels, finite mixture models, expectation-maximization, maximum-likelihood, parameter estimation, convergence rate

## 1. Introduction

The bulk of machine learning literature has historically focused on supervised learning, as a result of the mathematical tractability and wide range of favorable estimation properties it enjoys (Bishop, 2006). Yet, the preventive (often, prohibitive) cost of retrieving labeled datasets in numerous applications has raised an increasing interest in unsupervised learning approaches (Duda et al., 2001). The latter remains the sole theoretical tool at hand in practical scenarios, where the unavailability of either reward signals or even a limited set of labeled instances renders both reinforcement (Sutton and Barto, 1998) and semi-supervised learning methods (Chapelle et al., 2006) equally unsuitable. Recent works have showcased that, even in this setting, there exist ways to improve the quality of parameter estimation over conventional unsupervised techniques by exploiting additional, side-information on the model parameters. Such information is typically injected into the optimization problem in the form of assumed constraints or prior knowledge.

Along these lines, the present article formulates and studies algorithms that exploit contextual information to improve maximum-likelihood (ML) parameter estimation in generative finite mix-

ture models (FMM) (Bishop, 2006, chap. 9). More specifically, the proposed algorithms assume the possibility to extend the probabilistic directed graph of FMMs with contextual random variables  $c_i$  whose prior,  $p(c_i)$ , and/or conditional distributions,  $p(z_i|c_i)$  or  $p(c_i|z_i)$ <sup>1</sup>, are known, thus providing the additional information at the learner’s service. Context can be defined as any measurable entity having a known dependency relationship to the latent label, yet, is not part of the learning problem’s feature space but, rather, of its contextual environment. For instance, in a handwritten symbols classification problem with image-based features, a language model (context) contains valuable information on a symbol’s identity, which cannot be efficiently represented directly into the feature set.

The main motivation of this work is to derive algorithms able to learn “better” than their common unsupervised equivalents, and as close as possible to the supervised ones, while always respecting a strict restriction of non-explicit supervision (aka, absolutely no manual data label collection of any kind). A second important motivation is to provide simple, and intuitive derivations and formulations of such algorithms, in contrast to the majority of works in the relevant literature of learning with side-information (Section 2). In order to achieve the latter goals, the algorithms proposed here are limited to a specific framework yielding the following characteristics: Firstly, standard maximum-likelihood estimation (MLE) by virtue of the expectation-maximization (EM) method on extended FMMs. Since its formalization by Dempster et al. (1977), EM-MLE constitutes the cornerstone of unsupervised learning for probabilistic models and by far the most popular and understandable unsupervised method. Secondly, our algorithms are restricted to exploit probabilistic representations of contextual, side-information, which is, additionally, internal to the model. The (naturally, also probabilistic) model employed throughout this article is that of generative FMMs, chosen for being very generic in itself, while also yielding a simple, minimal probabilistic graph. A third motivation and novelty in our work is the provision of theoretical evidences explaining how contextual information yields parameter estimation benefits.

Following the aforementioned motivations, the contributions of this work are threefold. First, we introduce the analytic formulation of two contextually-enhanced EM-MLE algorithms for FMMs within the framework outlined above<sup>2</sup>. We demonstrate their theoretically sound derivation and, therefore, their non-heuristic, principled nature, while also highlighting important differences between them. Second, we present results on artificial datasets that confirm improvements in various FMM scenarios in terms of parameter estimation precision, standard errors, convergence rates, as well as how these translate into more accurate classification or regression quality. Our study also provides a comparative analysis of the proposed algorithms against each other and against the standard supervised and unsupervised MLE algorithms. Additionally, an exemplary application demonstrates the applicability and effectiveness of this approach to real-world problems. The third contribution entails the in-depth study of the underlying mechanisms through which these algorithms improve the common unsupervised problem. This is achieved, on one hand, by analyzing exemplary likelihood landscapes; on the other hand, by employing Fisher information and the *missing information principle* (Orchard and Woodbury, 1972), we prove the generalization of benefits to all types of FMMs, as well as the fact that the inherent missing information is alleviated by the contextual assistance proportionally to its information content. The proposed algorithms are thus shown to operate at least somewhat better than the regular unsupervised EM-MLE FMM estimator for context of non-zero entropy, and as good as the supervised estimator for rich context, for all the parameter estimation metrics studied. Hence, although all relevant works regard the exploitation of some sort of additional information for learning, this is the first time that information-theoretic insights are offered to explain the positive effects of side-information.

In spite of the simplicity and intuitiveness of the proposed framework, to our best knowledge, no in-depth study exists in the relevant literature. Yet, the outlined approach also suffers two main limitations. First, being specific to probabilistic modeling of context embedded into the learned

1. Variable  $z_i$  represents the latent class label of data sample  $x_i$ .

2. Each algorithm is pertinent to one of the two possible types of dependency between  $c_i$  and  $z_i$ .

model—unlike recently introduced methods for general, constraint-based unsupervised learning like Posterior Regularization (PR) and Generalized Expectation Criteria (GEC)—it does not qualify as an entirely generic method for improved unsupervised learning. Furthermore, it is not always the case that for any given estimation problem context and its prerequisite statistics can be readily available or easy to collect. Still, given the ability of FMMs to fit a large variety of problems in itself, the intuitiveness of probabilistic modeling of context and the fact that exploitation of context is increasingly addressed in many real-world situations, it can be claimed that a wide application spectrum can benefit from the aforementioned algorithmic advantages.

The remainder of this manuscript is organized as follows: Section 2 discusses the relevant literature and highlights its differences with the present work. Section 3 presents the proposed algorithms and their derivations, the employed theoretical tools for analysis and the evaluation methodology. Section 4 initially showcases the “modus operandi” of the proposed algorithms with theoretical results in specific examples and, subsequently, studies their parameter estimation properties in various FMM problems, as well as their benefits in a selected application. Finally, Section 5 discusses the proposed approach at the light of the extracted results.

## 2. Related work

Our work naturally falls under the semi-supervised learning framework in a broad sense (Chapelle et al., 2006), as context exploitation is a weak type of supervision. However, our algorithms depart from the classical semi-supervised literature and bootstrapping approaches (McCallum and Nigam, 1999; Dasgupta et al., 2002) in that absolutely no labeled data is required for parameter estimation. Unlike what the intuitions about general usage of context might suggest, the presented approaches are only loosely related to domain adaptation/transfer learning techniques (Pan and Yang, 2010), as our goal is to improve learning within a single contextual environment rather than generalize among two or more of those. Our algorithms are thus best categorized into the relatively new class of unsupervised learning methods exploiting side-information.

A great deal of related literature addresses various cases of weak supervision that can emerge in applications where, although some form of data labels is available, it does not fully comply with the assumptions of regular supervised learning. In this broad category one could identify a number of different situations. First, learning from partially or ambiguously labeled datasets, where each data sample is associated to many possible labels only one of which is correct (Cour et al., 2011; Chen et al., 2013), as well as multi-label, multi-annotator (crowd-sourcing) settings where all of the labels could be valid, potentially with different and time-varying reliability (Tsoumakas et al., 2010; Sun et al., 2010; Zhang and Obradovic, 2011; Audhkhasi and Narayanan, 2013; Lakshminarayanan and Whye Teh, 2013). Liu and Wang (2012); Sellamanickam et al. (2012) study partial-label problems where data labels are only missing for some of the classes. Multiple-instance or multi-view learning methods, where each learning example contains a bag of samples instead of a single one are studied by Foulds and Smyth (2011); Luo and Orabona (2010). Joulin and Bach (2012) propose a generic method to handle most of the above problems. Nguyen et al. (2011) put forward a framework exploiting additional information in the form of reliability indices of each data label. Similarly, Bouveyron and Girard (2009); Yasui et al. (2004); Uner et al. (2012) address cases with noisy or wrong labels. All the aforementioned approaches solve specific situations of weak supervision that differ from the setting discussed in our work, our main comparative advantage being that absolutely no manual labeling of any type is required, since the derivation of probabilistic labels through context in our scenarios is only implicit.

Another class of related problems are those where additional, side-information is provided in the form of constraints. Most of the early work in this category has focused on known positive and/or negative linkage between pairs or sets of samples (Basu et al., 2002; Shental et al., 2004; Georgi, 2009). These approaches were still tied to specific types of side-information. Given the latter observation, it is necessary to discuss methods that are able to cope with context-aware learning in its most wide

sense, irrespectively of whether side-information comes in the form of constraints, weak labeling or otherwise. Four main generic frameworks have been so far presented in the literature, differing basically in the exact way of representing the additional information and the objective function (as well as its optimization algorithm) adopted thereby to embed it into the learning problem.

Chang et al. (2007), later employed by Carlson et al. (2010), have proposed what the authors have called constraint-driven learning (CODL), which penalizes constraint violations of a given model by augmenting the objective function with a penalty term. This method has historically been the first principled approach towards generic weak learning. Nevertheless, its formulation assumes some labeled instances for model initialization, does not maintain uncertainty during learning, and involves a fairly heuristic optimization algorithm with a lot of hyperparameters.

Liang et al. (2009) put forward a Bayesian approach by modeling side-information as so-called “measurements”, defined as noisy expectations of constraint features. The employed objective function is optimized with a variational approximation assisted by further assumptions to induce mathematical tractability. This complex optimization procedure is the method’s main disadvantage. In a series of articles, McCallum and colleagues have introduced Generalized Expectation Criteria (GEC), where the additional information comes as linear constraints of a set of feature expectations forming a standalone objective or augmenting the common likelihood objective with an extra term (Mann and McCallum, 2010). A special case of GEC had been initially proposed as “expectation regularization” (Mann and McCallum, 2007). Based on this method, several optimization procedures have been presented and tested, including gradient descent (Druck et al., 2008) and variational approximation (Bellare et al., 2009). A special semi-supervised case of GEC has been independently formulated by Quadrianto et al. (2009).

Using the very same modeling of constraints, Ganchev et al. (2010) have proposed the Posterior Regularization (PR) framework, where constraints are imposed directly on the posterior distributions of latent models, giving rise to optimization algorithms akin to regular EM. The conceptual intuitiveness of this formulation for constrained-driven learning has led to many applications of PR presented thus far in the literature (Chen et al., 2011; He et al., 2013; Bryan and Mysore, 2013; Yang and Cardie, 2014; Zhu et al., 2014). Ghosh et al. (2009) have independently proposed a PR formulation specific to FMMs and constraints in the form of a-priori knowledge of mixing proportions, thereby deriving a variant of the “scaled”-PR algorithm for this particular problem (Ganchev et al., 2010, Appendix A). This work, despite sharing the same basic model of our algorithms, fails to see the benefits of probabilistically embedding context into FMMs and the different algorithmic possibilities generated thereby, while also resulting in a more complicated final objective.

In a brilliant analysis, Ganchev et al. (2010, Section 4) show that under certain approximations, all four aforementioned frameworks are in fact equivalent. In comparison to our work, it can be said that our algorithms trade-off a certain amount of genericity in favour of algorithmic simplicity and intuitiveness. This comes as a result of the fact that these frameworks employ constraint features that are external to the model, while in our cases context is embedded in the model itself. As a result, PR and GEC exhibit higher flexibility in representing diverse types of constraints, at the expense of more complicated optimization and modeling procedures. These claims are further substantiated in Appendix B, where the PR-equivalents of our algorithms are discussed.

To better appreciate the contributions of this paper, it is essential to elaborate on these works that are particularly similar to ours, either in the way of modeling context, or in the formulation of the derived algorithms. Starting with the first aspect, the idea of augmenting a given model to include context can be traced back to the “hierarchical shrinkage” method of McCallum and Nigam (1999). Both application-related constraints and probabilistic context modeling identical to ours are proposed by Kindermans et al. (2012a,b), who are additionally concerned with a brain-computer interface (BCI) application similar to the one we test in Section 4.4. However, in this case the authors focus on classification improvements rather than the estimation properties and analysis of the learning algorithm they propose. Lastly, a number of articles addressing the aforementioned “weak labeling” problems resort to probabilistic modeling of the additional labeling information (Cour

et al., 2011; Lakshminarayanan and Whye Teh, 2013), however, as already argued, such methods require explicit data labeling of some sort.

Moving to the second aspect concerning the formulation of the algorithm, two methods presented in the literature, following completely different avenues, have produced the same formulation of one of the algorithms proposed hereby, the one termed *WCA* (Weighted Context–Aware, Section 3.1). Bouveyron and Girard (2009) arrive to this derivation in the context of weak learning with noisy labels, where the extra weight expresses the probability of agreement between the potentially “noisy” labels and the data information. On the other hand, in what is probably the most relevant work to ours, Côme et al. (2009) derive the *WCA* algorithm assuming the existence of uncertain “soft” labels through Dempster–Shafer basic belief assignments, while also studying the effects of increasing contextual information as well as misspecified information. Our work is, first, more general than those, since this formulation of the problem coincides with one of the cases we study hereby. Second, our formulations are justified in a probabilistic setting, alternative to those employed in the aforementioned works, while also not requiring manual labels of any type. Most importantly, our scope is broader, since we provide theoretical insights on how the benefits of such frameworks can be explained in the context of the missing information principle, and study more deeply the parameter estimation and convergence rate properties.

### 3. Methods

In this section we describe the formulations and justify the derivations of the proposed algorithms for context-aware learning through probabilistic graphical models, briefly present the theory around the missing information principle and define the basic metrics and the methodology involved in our simulation studies presented in Section 4.

#### 3.1 Context–aware learning algorithms for FMMs

In order to gain a solid understanding of the proposed algorithms, the reader should recall (Bishop, 2006, chap. 9.2) that a conventional generative probabilistic FMM has the directed graph representation of Figure 1 (enclosed in dashed line), where  $\mathbf{x}_i \in X$  the observed independent and identically distributed (iid) data samples of a dataset  $X$  with cardinality  $N$  ( $i \in [1, N]$ ),  $\mathbf{z}_i \in Z$  the latent data representing the mixture<sup>3</sup> generating sample  $\mathbf{x}_i$  having a 1-of- $M$  representation<sup>4</sup>, so that  $z_{ij} \in \{0, 1\}$ ,  $\sum_j z_{ij} = 1$  and  $M$  the finite number of mixtures/classes. The distribution of observed data  $\mathbf{x}$  is then:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{j=1}^M \pi_j f_j(\mathbf{x}, \boldsymbol{\theta}_j) \quad (1)$$

where,  $\pi_j = p(z_j = 1)$  are the mixture coefficients with  $\sum_{j=1}^M \pi_j = 1$  and  $f(\mathbf{x}, \boldsymbol{\theta}') = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}')$  with  $f$  belonging to some identifiable parametric family with parameters  $\boldsymbol{\theta}'$ . ML estimation proceeds with maximizing the logarithm of the incomplete–data, marginal likelihood  $\log L(\boldsymbol{\theta}|X) = \log(\prod_{i=1}^N p(\mathbf{x}_i))$ . In supervised estimation,  $\mathbf{z}_i$  are instead observed (and referred to as the labels  $\mathbf{y}_i$ ) resulting in the marginal likelihood having (most often) a simple analytic solution. On the contrary, with latent  $\mathbf{z}_i$ , this optimization is intractable. However, the iterative, two–step EM–MLE can be employed, where instead of the marginal incomplete–data log–likelihood  $\log L(\boldsymbol{\theta}|X)$ , one first forms the expectation (under the posterior distribution of latent variables  $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$ ) of the complete–data log–likelihood  $\log L_c(\boldsymbol{\theta}|X, Z)$  (E–step):

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^k) = \mathbb{E}_{\hat{\boldsymbol{\theta}}^k} \{ \log L_c(\boldsymbol{\theta}|X, Z) \} = \sum_{i,j}^{N,M} \mathbb{E}_{\hat{\boldsymbol{\theta}}^k} \{ z_{ij} \} \log \pi_j + \sum_{i,j}^{N,M} \mathbb{E}_{\hat{\boldsymbol{\theta}}^k} \{ z_{ij} \} \log (f_j(\mathbf{x}_i, \boldsymbol{\theta}_j)) \quad (2)$$

3. The terms mixture, class and label will be used interchangeably hereafter.

4. Wherever notation is simplified, we adopt the alternative definition  $z_i \in [1, M]$  without warning.

where  $\theta = \{\pi_j, \theta_j\}, \forall j$  corresponds to the overall parameters to be estimated and  $\hat{\theta}^k$  their current estimate after  $k$  iterations of the algorithm; then,  $Q(\theta, \hat{\theta}^k)$  is analytically maximized (M-step):

$$\hat{\theta}^{k+1} = \underset{\theta}{\operatorname{argmax}}\{Q(\theta, \hat{\theta}^k)\} \quad (3)$$

This conventional unsupervised EM-MLE algorithm (termed hereafter *US*) is known to suffer certain limitations. More specifically, it tends to get stuck in local maxima of the marginal likelihood and, hence, is very sensitive to initialization of parameters  $\hat{\theta}^0$ , often producing estimates that are substantially different from those that would have been acquired in the supervised version of a given learning problem. Furthermore, the algorithm’s convergence rate (number of iterations until convergence) can be too slow for the requirements of certain applications, while also the standard errors are compromised compared to supervised estimation (termed *S* in the remainder of the manuscript).

It is clear that these limitations should be related, on one hand, to the missing label information. That is, since the nature of  $z_i$  (observed vs latent) is the only difference between supervised and unsupervised MLE of some FMM from a given dataset  $X$ . On the other hand, it is apparent that *US* only benefits from information coming from the data  $X$  and the current parameter estimates  $\hat{\theta}^k$  to retrieve knowledge on the latent data (*bottom-up* information), through computation of the posterior distribution  $p(z|\mathbf{x}, \hat{\theta}^k)$ . It can be hence assumed that improvement of the aforementioned estimation properties could arise by providing additional, independent from  $X$ , information on the latent class labels  $z_i$ . Unlike the approaches reviewed in Section 2, where such information is

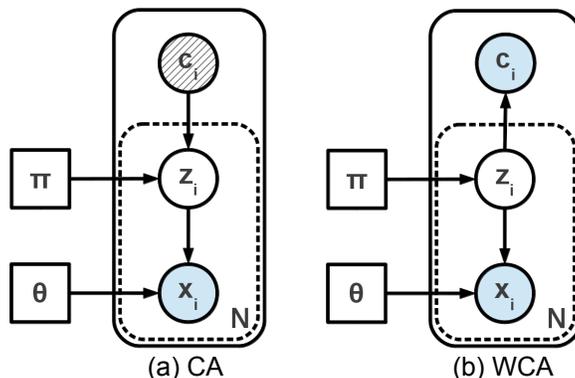


Figure 1: Graphical representations of augmented (solid boxes) and regular (dashed boxes) mixture models for a set of  $N$  independent and identically distributed (iid) data samples. Random variables depicted in circles, transparent for latent variables, shaded for observed variables and stripping for variables that can be observed or latent on occasion. Model parameters are illustrated with squares.  $\mathbf{x}_i$  are the observed data samples,  $z_i$  the latent class labels and  $c_i$  the contextual variables. Model (a) gives rise to *CA*-type of estimation and (b) to *WCA*.

retrieved either through direct but “weak” labels, constraints on or expectations of external features, our work studies the effects of additional information of probabilistic nature inherent to the model itself. These two principles set the basis of the subsequent formulation of the proposed algorithms.

First, respecting a probabilistic approach, the additional information is modeled as random variables  $c_i$ <sup>5</sup>. These are termed “contextual” variables since, from the application perspective and as introduced in Section 1, they are related to measurable contextual entities of the estimation

5. Without loss of generality, the contextual random variables will be assumed hereafter to be univariate and discrete in order to simplify the derivation of the algorithms.

problem’s environment that cannot be conventionally included as features by augmenting  $\mathbf{x}_i$ . Second, respecting the condition of the information being “internal” to the model, these variables are assumed to have a known dependence relationship to the latent labels  $\mathbf{z}_i$ .

This modeling has three desirable consequences; to begin with, the resulting augmented FMM models, shown in Figure 1, “encompass” the original FMM, thus leading to essentially the same estimation problem (of parameters  $\theta$ ) of the regular *US* algorithm. This is a direct result of the fact that the additional distributions involved in the augmented models,  $p(c_i)$  and/or  $p(\mathbf{z}_i|c_i)$ ,  $p(c_i|\mathbf{z}_i)$  are further assumed to be exactly known to the learner, and there is thus no additional parameters related to them to be estimated. The latter is a basic assumption which accounts for the extra, *top-down* information provided to the model compared to *US*. Albeit a strong one, it is equivalent to the assumption of known expectations of constraint features employed by GEC and PR, or the existence of weak labeling assumed by other methods. Furthermore, it complies with the intuition that one needs to have some knowledge of the contextual environment in order to exploit it, for learning or otherwise. Second, the augmented directed probabilistic graph representations imply that the very same simple, well-known and understood EM–MLE estimator used for the non-augmented graph and resulting in the *US* algorithm can be again employed. Third, the two possible types of dependence between  $c_i$  and  $\mathbf{z}_i$  give rise to two different augmented models (Figures 1a and b) and, therefore, two similar but different EM–MLE algorithms that can be used.

Given the above probabilistic framework, the derivation of the novel context-aware EM–MLE algorithms proceeds equivalently to the *US* case for the augmented graphical models. The objective function to be optimized is again the incomplete-data log-likelihood  $\log L$ , corresponding either to the marginal distribution  $p(X)$  or to  $p(X|C)$ . Both these cases are derived from the joint distribution of each augmented model,  $p(X, Z, C)$ , after conventionally marginalizing out variables  $Z$ . Additionally, the contextual variables  $C$  should be either also marginalized out or conditioned upon, respectively. The intermediate objective, namely, the expected complete-data log-likelihood  $Q(\theta, \hat{\theta}^k) = \mathbb{E}_{\hat{\theta}^k} \{\log L_c(\theta|X, Z)\}$  is similarly derived by  $p(X, Z)$  with marginalized  $C$  or  $p(X, Z|C)$ , under the posteriors  $p(Z|X, \hat{\theta}^k)$  or  $p(Z|X, C, \hat{\theta}^k)$ , respectively. Marginalizing  $C$  out is applied when the contextual assistance needs not be observed, while conditioning is mandatory in the opposite case. Table 1 summarizes the naming convention, E-step and  $\log L$  formulation of each algorithm considered here, while the analytical derivations of the algorithms along these lines are provided in Appendix A. The issues that need special attention are elaborated in the following paragraphs.

First, in comparison to *US*, both context-aware algorithms we propose alter only slightly the EM algorithm’s E-step, leaving the M-step unaffected, largely accounting for their ease-of-use and intuitiveness. Furthermore, the context-related terms of the quantities in Table 1, representing what has been called hereby the *top-down* information, can be isolated from the regular *bottom-up* information to implicitly define sample-wise probabilistic labels  $p_{ij}$  with  $\sum_{j=1}^M p_{ij} = 1$  (i.e., each  $\mathbf{p}_i$  is a discrete probability distribution over the latent variables  $\mathbf{z}_i$ ). Therefore, the entropy of these labels represents a measure of the contextual information content individually for each sample  $\mathbf{x}_i$  and, by averaging, for the overall estimation problem. Our work is thus the only one together with Côme et al. (2009), highlighting that the additional side-information is measurable and can be used to predict expected estimation benefits beforehand. The formulations of implicit probabilistic labels through context for each algorithm are illustrated in Table 1.

The outlined methodology for context-aware learning results in two similar algorithms, each of which is associated to one of the graphs in Figure 1, whose difference is the reversed dependency between variables  $c_i$  and  $\mathbf{z}_i$ . Dependency as in Figure 1a results in an E-step where, compared to the regular *US* case, the implicit probabilistic labels replace priors  $\pi_j$  in all the involved quantities:  $\log L$ ,  $\log L_c$  and E-step. This algorithm is hereafter termed as *CA* (Context-Aware). On the other hand, dependency as in Figure 1b leads to the *WCA* (Weighted Context-Aware) algorithm, where the probabilistic labels appear as additional terms in the aforementioned quantities and act as

---

6. Latent context  $C$ .

Table 1: Algorithms for maximum-likelihood estimation of FMMs. *US*: Regular unsupervised EM-learning. *CA*: Context-aware EM-learning. *WCA*: Weighted context-aware EM-learning. *DCA*: Direct context-aware learning. *S*: Regular supervised learning.

Algorithm	$\mathbf{p}_i$	E-step $\mathbb{E}_{\hat{\theta}}\{z_{ij}\} = \dots$	$\log L = \sum_{i=1}^N \log(\sum_{j=1}^M (\dots))$
<i>US</i>	None	$\frac{\pi_j f_j(\mathbf{x}_i   \hat{\theta}_j)}{\sum_{m=1}^M \pi_m f_m(\mathbf{x}_i   \hat{\theta}_m)}$	$\pi_j f_j(\mathbf{x}_i   \hat{\theta})$
<i>CA</i>	$\begin{cases} \sum_{c_i} p(c_i) p(\mathbf{z}_i   c_i)^6 \\ p(\mathbf{z}_i   c_i) \end{cases}$	$\frac{p_{ij} f_j(\mathbf{x}_i   \hat{\theta}_j)}{\sum_{m=1}^M p_{im} f_m(\mathbf{x}_i   \hat{\theta}_m)}$	$p_{ij} f_j(\mathbf{x}_i   \hat{\theta})$
<i>WCA</i>	$\frac{p(c_i   \mathbf{z}_i)}{p(c_i)}$	$\frac{p_{ij} \pi_j f_j(\mathbf{x}_i   \hat{\theta}_j)}{\sum_{m=1}^M p_{im} \pi_m f_m(\mathbf{x}_i   \hat{\theta}_m)}$	$p_{ij} \pi_j f_j(\mathbf{x}_i   \hat{\theta})$
<i>DCA</i>	Custom	$p_{ij}$	$\pi_j f_j(\mathbf{x}_i   \hat{\theta})$
<i>S</i>	$y_i$	$\begin{cases} 1 & , y_i = j \\ 0 & , y_i \neq j \end{cases}$	$\pi_j f_j(\mathbf{x}_i   \hat{\theta})$

“weights” of the conventional E-step of the *US* algorithm. Note that, as mentioned in Section 2, the formulation of the *WCA* algorithm is identical to that proposed by Côme et al. (2009); Bouveyron and Girard (2009), albeit thereby derived from different settings and requiring explicit labelers. It is further worth to note that *CA* “supports” both observed and latent context  $C$ , with identical formulation and slight modification in the definition of probabilistic labels  $p_{ij}$ , while the *WCA* algorithm necessitates observed context, in what can be thought of as an important advantage of the former algorithm. Further differences between the proposed algorithms are showcased in Section 4.

Inspired by the implicit derivation of probabilistic labels through context, we are also considering another algorithm termed *DCA* (Direct Context-Aware), where the posterior distribution is fully defined by such probabilistic labels alone. Unlike *US*, which derives this distribution solely from *bottom-up* information and algorithms *CA* and *WCA*, which fuse both *bottom-up* and *top-down* information, *DCA* employs only the latter. As such, *DCA* maximizes directly the complete-data likelihood  $\log L_c$  in a single iteration, just like the supervised ML estimator *S*. It only differs from the latter in that, instead of crisp, certain labels  $\mathbf{y}_i$ , probabilistic labels  $\mathbf{p}_i$  are employed. From this point of view, *DCA* can be thought of as a possibility for MLE of FMMs in the weak labeling setting discussed in Section 2, albeit as it will be shown in Section 4 it suffers significant drawbacks.

The central idea behind the algorithms introduced in our study is evident in the E-steps of Table 1. It is easy to see that, in the extreme case where the implicit probabilistic labels  $\mathbf{p}_i$  obtain the lowest possible entropy (aka, maximum information content), both the  $\mathbf{p}_i$ -s and the overall E-steps of all algorithms become identical to crisp labels  $\mathbf{y}_i$  of the regular supervised estimator *S*. In this case, all algorithms will yield the same supervised MLE through the M-step, because of the profound similarities of the expected complete-data log-likelihoods  $\log L_c$  in Equations 2, 22, 24. It is hence reasonable to foresee that even uncertain contextual assistance, whose information content is however greater than the minimum, will bias somewhat the MLE towards that of supervised estimation. Obviously, the latter forms an upper bound of the overall information that can be available in a given estimation problem. That being said, our context-aware algorithms are able to bring forward these advantages without requiring any actual label collection, just like the regular

unsupervised algorithm *US*; the proposed algorithms are thus compromising and making the best out of these two worlds.

Last but not least, since practically our algorithms only require mild modifications on the regular *US* algorithm’s E–step, one might be tempted to consider them as simplistic heuristics. Yet, as discussed above and shown in Appendix A, they are in fact direct consequences of the probabilistic modeling and embedding of contextual information adopted here. This claim is solidly supported by considering the fact that, what each algorithm is really modifying, are the initial and intermediate objective functions maximized,  $\log L$  and  $\log L_c$ , respectively (in a way that those depart from the *US* objective and approach that of *S*). The modification of the E–step is only a natural consequence of the latter, and thus not heuristic, while simplicity remains a clear advantage. The proof of convergence for the *CA* algorithm is provided in Appendix A, following the standard reasoning employed for the *US* algorithm. The equivalent proof along the same lines for *WCA* is skipped, but can be found in Côme et al. (2009). Finally, Appendix B elaborates on the relation of the proposed framework to that of Posterior Regularization (PR), showcasing that, despite seemingly interchangeable, our own derivation maintains certain clear advantages over PR.

### 3.2 Information matrices and missing information principle

As already discussed in the previous subsection, intuition suggests that any parameter estimation improvement brought forward by context–aware MLE should be attributed to the additional contextual information at hand, compared to the conventional, unsupervised MLE. Nevertheless, in order to shed further light on this fundamental issue, one has no better option but to rely on the Fisher Information (Lehmann and Casella, 1998), the most formal and well-studied way of measuring the amount of information involved in the estimation of the unknown parameters  $\theta$ . Therefore, we set out to study, for each algorithm considered here, approximations of the (*expected*) Fisher information matrix  $I(\theta)$  through its sampled-based version, the *observed information matrix*  $I(\theta|X)$ . The latter measures the amount of information a sample  $X$  carries on the estimated parameters  $\theta$ , where  $I(\theta) = \mathbb{E}_{\hat{\theta}}[I(\theta|X)]$  and  $I(\theta|X) = -\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta^T} \Big|_{\theta = \hat{\theta}_{ML}}$ , the negative of the Hessian of the log-likelihood objective function evaluated at the ML estimate.

Orchard and Woodbury (1972) proved that the observed information for missing–data problems can be computed as the difference  $I(\theta|X) = I_c(\theta|X) - I_m(\theta|X)$ . The first term is the conditional expectation of the complete–data information matrix given the observed data, an estimate of the available information if there were no missing data. The second term, called the *missing information matrix*, is the expected information for  $\theta$  based on the missing data  $Z$  when conditioned on observed data  $X$ , representing the information lost due to missing data. This relation has been called the *missing information principle* (MIP).

Both these matrices can be computed through complete–data quantities (so that their calculation is tractable), as:

$$I_c(\theta|X) = \mathbb{E}_{\hat{\theta}} \left\{ -\frac{\partial^2 \log L_c}{\partial \theta \partial \theta^T} \right\} \Big|_{\theta = \hat{\theta}_{ML}} \quad (4)$$

$$I_m(\theta|X) = \text{cov}_{\hat{\theta}} \{ \mathbf{S}_c(X|\theta) \mathbf{S}_c(X|\theta)^T \} \Big|_{\theta = \hat{\theta}_{ML}} \quad (5)$$

where,  $\mathbf{S}_c(X|\theta)$  is the score (gradient vector) of the complete–data log-likelihood. The application of the missing information principle on our algorithms, studied in conjunction with the measurable amount of contextual information (entropy of the implicit probabilistic labels) becomes the instrument to prove formally the ability of the proposed algorithms to alleviate the missing information in unsupervised FMM estimation (Section 3.2 and Appendix C) for all types of FMMs. It should be underlined that, although all literature presented in Section 2 is concerned with the problem of exploiting side–information to improve unsupervised estimators, our work is the first to unveil and explain the acquired estimation benefits using information–theoretic principles.

Studying the Fisher information in our algorithms also proves to be advantageous for estimating benefits in terms of additional estimation metrics, beyond estimation precision. More specifically, the observed information allows for the computation of the variance–covariance matrix of the MLE, as  $C = I^{-1}(\boldsymbol{\theta}|X)$  and, hence, the standard errors of parameter estimation as  $SE_i = \sqrt{I_{i,i}^{-1}(\boldsymbol{\theta}|X)}$  for the  $i^{\text{th}}$  parameter in vector  $\boldsymbol{\theta}$ . The standard errors associated to our algorithms for different amounts of contextual information can be thus derived painlessly, without resorting to repeated sampling.

Furthermore, an index of the algorithms’ convergence rates can be derived and compared through the matrices involved in the MIP, as follows: when EM converges to a local maximum, it has been shown (Dempster et al., 1977) that the convergence rate  $r = \lim_{k \rightarrow \infty} \left\| \frac{\hat{\boldsymbol{\theta}}^{k+1} - \hat{\boldsymbol{\theta}}^k}{\hat{\boldsymbol{\theta}}^k - \hat{\boldsymbol{\theta}}^{k-1}} \right\|$  is linear. It further coincides with the spectral radius ( $\lambda_{max}$ , where  $\lambda_i \in [0, 1), \forall i$ , the eigenvalues) of the “rate” matrix  $J$ , defined as  $J(\boldsymbol{\theta}) = I_c^{-1}(X|\boldsymbol{\theta})I_m(X|\boldsymbol{\theta})$  and expressing the total fraction of missing information (McLachlan and Krishnan, 2008). Convergence rate is therefore also related counter–proportionally to the amount of missing information and, hence, expected to improve by our context–aware algorithms. In Section 4 we use the definition  $r' = 1 - r = 1 - \lambda_{max}$  to measure the theoretical rate of convergence. This definition complies with the intuition that 0 corresponds to a non-converging algorithm and 1 to an algorithm that converges immediately, in a single iteration.

### 3.3 Evaluation metrics and simulation design

In Section 4 we perform a comparative analysis of the proposed algorithms, *CA*, *WCA* and *DCA*, against the regular supervised and unsupervised estimators, *S* and *US*, respectively. The comparison entails, first, artificially constructed datasets and, second, real–world data from a selected application. For each considered scenario, our results are extracted for different amounts of contextual information content. That is to stress the fact that, in our case, the “strength” of side–information is measurable and, most importantly, the estimation benefits depend on it.

The estimation properties reported are *estimation precision*, *standard errors* and *convergence rate*. Estimation precision is evaluated by means of the Euclidean distance between the estimated parameter vector  $\hat{\boldsymbol{\theta}}$  and the actual one  $\boldsymbol{\theta}^A$ , namely:  $D = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^A\|$ . Concerning standard errors, we employ the theoretical estimation of standard errors  $SE_i$  derived through the information matrices (see Section 3.2). For brevity, we report only the average of the standard errors of the estimated parameters  $ASE = \frac{1}{L} \sum_{i=1}^L SE_i$  (where  $L$  is the number of parameters estimated). Similarly, regarding the convergence rate, the theoretical estimates of one minus the fraction of missing information,  $r' = 1 - r$ , is used. The classification performance of trained models is assessed through N-class classification accuracy  $A = N_c/N$ , where  $N_c$  the number of correctly classified samples out of  $N$  total samples across all classes. For problems with unbalanced number of samples per class, we employ instead a “balanced” accuracy metric *BA*, which is simply the arithmetic mean of class–wise recalls, which is more objective in this situation. Note that  $A$  and *BA* coincide for balanced problems. Finally, for regression tasks, the mean square error (MSE) is reported.

As already motivated, we wish to test the performance of the presented algorithms at different levels of contextual information content that might be available, which accounts for one of the strengths of our work in comparison to the literature (only adopted by Côme et al. (2009)). The implicit extraction of probabilistic labels  $\mathbf{p}_i$  for each sample  $\mathbf{x}_i$  through context offers the opportunity to directly use the common entropy metric. Nevertheless, in order to uniformize our information metric across problems with different number of mixtures/classes, conveniently bound it between  $[0, 1]$  and comply with the intuition that low metric value corresponds to low information content, we employ instead a scaled negentropy definition:  $NE_i = 1 + \sum_{j=1}^M p_{ij} \log_M p_{ij}$ ,  $NE_i \in [0, 1]$ . This metric evaluates the information of each probabilistic label individually; as an index for an overall dataset, the average across all labels can be employed. It is easy to see that  $NE_i$  will assume its lowest value, 0, irrespectively of the number of mixtures  $M$ , when  $\mathbf{p}_i$  is uniform,  $p_{ij} = 1/M, \forall j \in [1, M]$ ; in other words, when the contextual information does not cast a preference over some class, a situation

which will be called “ignorant” context. Conversely,  $NE_i$  will be 1 when  $p_{im} = 1, m \in [1, M]$  and  $p_{ij} = 0, \forall j \neq m, j \in [1, M]$ ; essentially, this is the case when context is “perfect”, fully revealing the latent class label  $\mathbf{z}_i$ , since  $\mathbf{p}_i$  will be “crisp” and identical to explicit labels  $\mathbf{y}_i$  provided in the supervised setting.

While in a real application the probabilistic labels  $\mathbf{p}_i$  are derived as shown in Table 1, for the simulation studies on artificial datasets,  $\mathbf{p}_i$ -s are directly provided. A label  $\mathbf{p}_i$  for sample  $\mathbf{x}_i$  is constructed randomly, so that its information content is exactly  $NE_i$ . For all but the “mixed” context scenario (see below), all samples of a generated dataset  $X$  are assigned the same  $NE$  value ( $NE_i = NE, \forall i$ ), so that this value can reflect exactly the contextual information content of the overall dataset for analysis purposes. Obviously, multiple probabilistic labels of the same  $NE$  exist. For instance, both labels  $\mathbf{p}_i = [0.757 \ 0.243]$  and  $\mathbf{p}_i = [0.243 \ 0.757]$  correspond to  $NE_i = 0.2$  in a 2-class problem, irrespectively of whether the ground-truth label is  $\mathbf{y}_i = [0 \ 1]$  or  $\mathbf{y}_i = [1 \ 0]$ . In all but the “wrong” context scenario (see below),  $\mathbf{p}_i$ -s are constructed to cast greater confidence to the ground-truth label  $\mathbf{y}_i$  (“correct” context). Formally, we impose that  $\arg \max\{\mathbf{p}_i\} = \arg \max\{\mathbf{y}_i\}$ , so that context always “predicts” the correct true label with increasing confidence as  $NE$  increases. This rule is only abandoned in the “wrong” context scenario, where the effects of misleading contextual information are investigated. In this scenario,  $k_i = \arg \max\{\mathbf{p}_i\} \neq \arg \max\{\mathbf{y}_i\}, k_i \in [1, M] \forall i$  is selected randomly out of the  $M - 1$  remaining possibilities for a percentage of the generated  $\mathbf{p}_i$ -s.

In Appendix C, it is formally proved that the missing information principle applies for any type of (identifiable) mixtures, number of mixtures, number of estimated parameters and irrespectively of whether an FMM estimation problem is univariate or multivariate. Still, our simulation results on artificial datasets show the estimation benefits (and their magnitude) in practice. Consequently, a series of FMM estimation problems are solved for all examined algorithms keeping the statistics of all aforementioned metrics for the following scenarios: (a) a mixture of two univariate normal distributions, where variances are known and only the two class means are estimated, (b) a mixture of two univariate normal distributions, where all existing parameters are estimated, (c) a mixture of three univariate normal distributions, (d) a mixture of two multivariate (2-dimensional) normal distributions, (e) a mixture of two univariate Maxwell–Boltzmann distributions and, finally, (f) a mixture of two univariate, first order linear regressors. In all the above scenarios, unless otherwise specified, all parameters  $\boldsymbol{\theta}$  are estimated.

Another three scenarios are considered for evaluating special situations related to the provided contextual information or the problem’s structure. More specifically, two scenarios study the effects of “mixed” and “wrong” context as already documented above, and a third one investigates the performance of context-aware algorithms in situations with unbalanced number of samples per class (“biased” scenario). All these scenarios employ a mixture of two univariate normal distributions, where all existing parameters are estimated.

For each of the above scenarios, 1000 estimation problems are generated and solved for all compared algorithms. Each problem  $r \in [1, 1000]$  is associated to a randomly generated dataset  $X_r, Y_r, P_r^{NE}$  of observed data, ground-truth labels and probabilistic labels, respectively. For the proposed algorithms *CA*, *WCA* and *DCA*, each problem  $r$  is further solved for different contextual information content  $NE \in [0 : 0.1 : 0.99]$  (generated as detailed above), so that our evaluation encompasses the complete range of possible contextual information content.  $NE = 1$  is not tested as it has been already shown to yield the supervised estimator  $S$  for all the context-aware algorithms. The cardinality  $N$  of each dataset is fixed in all cases to 100 times the number of parameters to be estimated, a rule of thumb that is known to produce sufficient data for the regular *US* algorithm. The ground-truth  $Y_r$  is constructed to have balanced number of samples per class, with the exception of the aforementioned “biased” scenario.

The observed data  $X_r$  are randomly generated from semi-randomly selected “actual” distributions. More specifically, all parameters in  $\boldsymbol{\theta}_r^A$  but one, are drawn from a uniform distribution each, whose boundaries are reported in Appendix D for each scenario. The remaining parameter is analytically computed so that a given problem  $r$  corresponds to a certain separability level  $SKL_r$  between

the involved mixtures, as quantified through Kullback–Leibler divergence.  $SKL$  is randomly selected for each problem and mixture from another uniform distribution, whose boundaries impose a range of very overlapping to very (but never completely) separable problems (see Appendix D). Any effect of separability on parameter estimation thus vanishes by averaging the results across the 1000 problems solved for each scenario. The initial parameter vector  $\hat{\theta}_0$  (common to all conditions/algorithms tested for a particular problem) is drawn from a similar procedure, where the Kullback–Leibler separability  $IKL$  between the actual and the initial  $j^{th}$  mixture is again drawn randomly from a uniform distribution (see Appendix D). This procedure maintains randomness in initialization (the most common technique in the absence of any prior information), while also ensuring that any illustrated effects on parameter estimation related to initialization can be ruled out, by averaging across 1000 problems. Mixing coefficients  $\pi_j$  are excluded from the above procedures. For all but the “biased” scenario, their actual and initial values are set to  $\pi_j^A = \hat{\pi}_j^0 = 1/M$ , resulting in “balanced” estimation problems. Initial mixing coefficient distributions are also set to be uniform in all scenarios, including the “biased” one, as per regular convention. All algorithms are left to perform as many iterations  $t$ , as needed so that  $\|\hat{\theta}_r^t - \hat{\theta}_r^{t-1}\| < 10^{-5}$ . If this stopping criterion is not reached after 300 iterations for some of the tested algorithms, estimation is stopped and  $\hat{\theta}_r^{300}$  is used as the final estimate.

We finally report for each scenario the average and standard deviation of the defined estimation metrics across all problems: estimation precision as  $D$ , average of standard errors  $ASE$  and convergence rate  $r'$ . In addition to that, average and standard deviation of classification accuracy  $A$  on the same scenarios are also illustrated (equivalently,  $MSE$  for the mixture of regressions scenario). Accuracy is computed for each scenario, problem  $r$  and algorithm, by generating a second “testing” dataset  $X'_r, Y'_r$  (of equal cardinality to  $X_r$ ) from the same “actual” FMM, which is classified using the estimated parameters of each algorithm by means of the Maximum–A–Posteriori rule. For the mixture–of–regressions scenario, the same evaluation methodology is applied to derive the  $MSE$  on the testing set. Wilcoxon ranksum tests are used to demonstrate eventual statistically significant differences of the proposed algorithms against  $US$  and  $S$ , as well as among each other for the same  $NE$  levels. For all metrics, significance at a 99% ( $\alpha = 0.01$ ) confidence interval is extracted.

## 4. Results

In this section, we first provide theoretical insights on how the proposed context–aware algorithms improve FMM parameter estimation and establish the applicability of the missing information principle. We then present estimation precision, standard error, convergence rate and resulting accuracy or Mean Square Error (MSE) results in various FMM parameter estimation scenarios as explained in Section 3.3. Finally, classification results with the  $CA$  algorithm are presented in an online–learning problem in brain–computer interaction.

### 4.1 Likelihood landscapes

In order to intuitively demonstrate the effects of the proposed algorithms, Figure 2a illustrates the incomplete–data log–likelihoods  $\log L(\mu_1)$  and the intermediate objectives maximized at the first M–step (complete–data log–likelihood expectations plus the latent data entropy)  $Q(\mu_1, \hat{\mu}_1^0) + H(\hat{\mu}_1^0, \hat{\mu}_1^0)$  (see Appendix A) for a mixture of two univariate Gaussians problem, where only  $\mu_1$  is estimated starting from  $\hat{\mu}_1^0 = 2$  and the rest of the parameters are fixed to  $\pi_1 = 0.1, \mu_2 = 1, s_1 = 0.5, s_2 = 3$ . Estimation is based on  $N = 100$  samples randomly generated from the above distribution using  $\mu_1 = 0$  (true parameter value). The above artificial setting is selected because it leads to multiple local maxima (when  $N$  is small) with a maximum close to the true parameter value (which is not necessarily the global maximum), so that the effects of the proposed algorithms can be clearly illustrated.

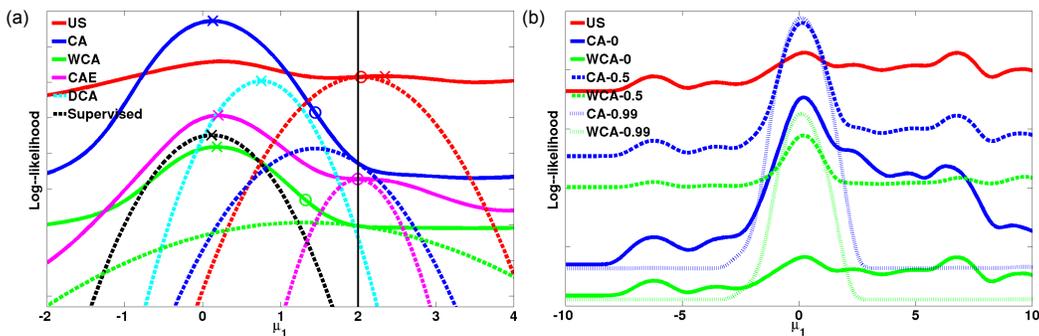


Figure 2: (a)  $\log L(\mu_1)$  (solid lines) and  $Q(\mu_1, \hat{\mu}_1^0) + H(\hat{\mu}_1^0, \hat{\mu}_1^0)$  (at first EM iteration, dashed lines) for various estimation algorithms as color-coded in the legend, in a mixture of two univariate Gaussians model. *CAE* refers to algorithm *CA* with ignorant context and *CA*, *WCA* and *DCA* are shown for contextual negentropy  $NE = 0.7$ . ML estimates after convergence shown with ‘x’ and estimates after the first iteration in ‘o’. (b) Incomplete-data log-likelihoods  $\log L(\mu_1)$  in the same problem as in (a) for *US* (red solid) and various contextual negentropy levels as shown in the legend for *CA* (blue) and *WCA* (green).

The first thing to note is that the intermediate objectives  $Q(\mu_1, \hat{\mu}_1^0) + H(\hat{\mu}_1^0, \hat{\mu}_1^0)$  (the results of the first E-step of each method), like EM theory dictates, are tangent to the respective complete-data log-likelihoods  $\log L(\mu_1)$  at the initial point  $\hat{\mu}_1^0 = 2$ . They also have the same gradient as the corresponding  $\log L(\mu_1)$ , and, thus, they form a lower bound of  $\log L(\mu_1)$ , which is maximized at the first M-step to yield new estimates  $\hat{\mu}_1^1$  (circles ‘o’ in Figure 2a). Besides illustrating the sanity of the methods in the EM setting, it is also shown that against initial intuition, context-aware methods do not simply alter the E-step formulation; the latter is actually the result of forming a new incomplete-data log-likelihood  $\log L(\theta)$ , different from that of the “regular” *US* method, as shown in Table 1.

The benefits of these methods (in case of “correct” context as in the example) come as a result of the features of the contextually modified  $\log L(\theta)$ . Besides a vertical translation that is unimportant for parameter estimation, the modified log-likelihoods (here with  $NE = 0.7$ ) tend to have a larger maximum close to the supervised estimate (black ‘x’); hence, also closer to the true parameter value. This maximum will thus also tend to be the global maximum, while the other local maxima are suppressed. The magnitude of these effects increases with increasing contextual negentropy  $NE$ , as shown in Figure 2b, where as  $NE \rightarrow 1$  (dotted lines) the problem reduces to supervised learning. In our example, this suppression allows all context-aware methods to “escape”, in contrast to the *US* method, the local maximum of *US* on the right side of the initial value and converge to the first local maximum on the left side of the initial value (signified with ‘x’ of the respective color). Since the latter is (naturally) much closer to the true parameter value, higher estimation precision and less sensitivity to initialization (two points where *US* is known to perform poorly) are achieved.

It is straightforward to assume that this favourable distortion of contextually enhanced  $\log L$ -s (comparative to that of *US*) should follow from the redistribution of the soft mixture assignments for each sample achieved in favour of the “correct” mixture thanks to the probabilistic labels. This hypothesis also explains the proportional dependence of effects on  $NE$ . This refined distribution of confidence in the definitions of both  $\log L$  and the E-step is thus the basic operational principle of context-aware learning algorithms. The observed effects should generalize for any type of FMM, as the confidence redistribution in the posteriors is independent of the number or type of mixtures.

The improvement of convergence rate and its dependence on contextual negentropy are also implied in the example, since the estimates  $\hat{\mu}_1^1$  for *CA* (blue ‘o’) and *WCA* (green ‘o’) are much closer

to their final MLE than for *US* (red) or *CAE* (*CA* with ignorant context, magenta), already after the first iteration. This fact is further substantiated in the next subsection. It is also interesting to note that algorithm *DCA*, though it still yields a better estimate than *US* (at least with the initialization  $\hat{\mu}_1^0 = 2$  selected for this example), results in a larger bias than the other context-aware algorithms, suggesting that discarding bottom-up information is suboptimal. Finally, in the case of ignorant context ( $NE = 0$ ), *WCA* (solid green line, Figure 2b) reduces to a translated version of *US* (red), while *CA* (blue) already “boosts” the favorable maximum. Hence, as will be also verified later, in the “ignorant” context case, *WCA* is identical to *US* (what can be also analytically shown by the E-step definitions in Table 1), while *CA* can already yield some improvement over *US*.

#### 4.2 Missing information principle, standard errors and convergence rate

The application of the *missing information principle* (MIP) on algorithms *CA*, *WCA* and *US* is demonstrated in a mixture of two univariate Gaussians problem with  $\pi_1 = 0.6, \mu_1 = 0, \mu_2 = 1, s_1 = 1, s_2 = 2$ , where standard deviations are fixed and  $\pi_1, \mu_1, \mu_2$  are estimated from initial guesses  $\hat{\pi}_1^0 = 0.5, \hat{\mu}_1^0 = 0.49, \hat{\mu}_2^0 = 0.51$ . For increasing values of contextual negentropy  $NE$ , we perform 100 repetitions randomly generating  $N = 10^4$  samples from the above distribution and estimate the standard errors of parameters  $\pi_1, \mu_1, \mu_2$  and the algorithms’ convergence rate, as theoretically predicted by means of estimating the corresponding information and rate matrices for each algorithm (see Section 3.2). For a certain contextual negentropy value, a final standard error for each parameter and a convergence rate value is derived by averaging across the 100 repetitions.

The analysis of the MIP for the proposed algorithms is necessary for a number of reasons. First, for algorithms investing in additional, side-information, it is essential to explain the acquired benefits from an information-theoretic perspective. That is, so that a theoretical link between the additional information and the one finally provided for parameter estimation (i.e., the observed information matrix  $I$ ) can be established. Surprisingly, despite broad relevant literature on the topic, our work is the first to provide such a result. Second, as outlined in Section 3.2, the MIP provides the tools to directly assess important (but, so far, largely overlooked) estimation properties, like standard errors and convergence rate. Even for metrics like estimation precision (what also affects subsequent classification accuracy and/or regression quality), towards which a direct link with the available information cannot be established (that is why, in the previous subsection, we opt for an explanation of effects based on likelihood landscapes), the MIP offers an indirect explanatory mechanism. Third, the MIP allows to explicitly show that the available information in context-aware MLE will be bounded between the information available in the supervised and unsupervised versions of a given estimation problem; hence, prospective users of the proposed algorithms can be aware of the best- and worse-case scenarios that can occur. Related to that, it is shown that manipulating the available information through context is not fully defined by the formulation of the algorithms per se, but also proportionally dependent to the information content of side-information. Last but not least, the MIP highlights important differences between the proposed algorithms.

These points are elaborated below, using the aforementioned example. It should be noted that this exemplary problem is randomly selected among infinite possibilities for illustrating the concept and effects of the MIP. The analysis in Appendix C and the results across 1000 problems in different scenarios in Section 4.3 show that the effects found here are generalizable to all FMM estimation problems.

For one repetition of the aforementioned problem, Figures 3a (for *WCA*) and 3b (for *CA*) show that matrices  $I_c$  (first row) remain unaffected by increasing contextual assistance. That is reasonable, since the complete-data information  $I_c = I^S$  (an estimate of the information available in the supervised setting) should be independent of any additional information. Additional information on the data labels is irrelevant to the complete-data statistics, which assume labels to be known. On the contrary, the magnitudes of the elements of the missing information matrix  $I_m$  (second row), which for  $NE = 0$  obtains its maximum  $I_m^{MAX} = I_m^{US}$  (when no additional information on missing

labels exists), are reduced with increasing contextual negentropy. They eventually vanish into the 0 matrix as  $NE \rightarrow 1$  ( $I_m^{MIN} = I_m^S = 0$ , since data labels are known or fully revealed by context, and there is no missing information associated to them). Consequently, the fractions of missing information (rate matrix  $J = I_c^{-1} * I_m$ , third row) also vanish, along with its spectral radius, expressing the total fraction of missing information.

Overall, as motivated, the missing label information in context-aware EM learning is shown to be eliminated to a certain degree, proportionally to the information content of the additional side-information. The finally available information in these algorithms, as encoded in the information matrices  $I = I_c - I_m$  (MIP definition), will consequently be bounded: Above, by  $I^{MAX} = I_c^{MAX} - I_m^{MIN} = I_c^S - 0 = I^S$  (identical to the supervised MLE) and below by,  $I^{MIN} = I_c^{MIN} - I_m^{MAX} = I_c^S - I_m^{US} = I^{US}$  (identical to the unsupervised MLE)<sup>7</sup>.

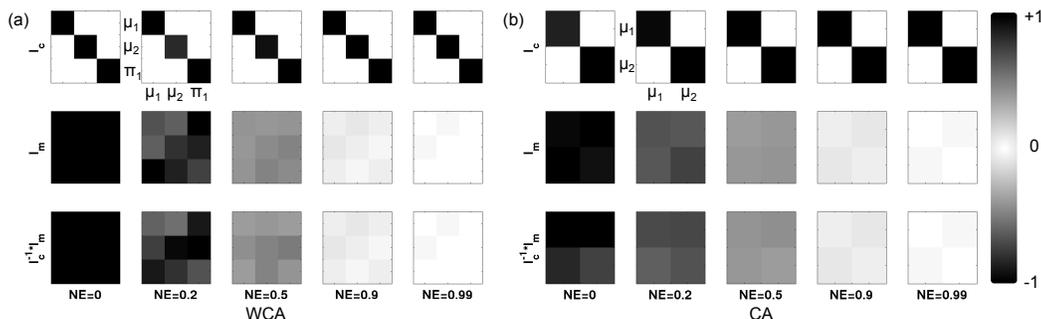


Figure 3: Information matrices  $I_c$ ,  $I_m$  and rate matrix  $J = I_c^{-1} * I_m$  with increasing contextual negentropy  $NE$  in a mixture of two univariate Gaussians estimation problem (see text), for (a) *WCA* and (b) *CA*. All matrix values normalized to  $[-1, 1]$  by dividing with the respective absolute value of the *US* algorithm.

The analytic dependences of standard errors on the information matrix  $I$  and of convergence rate on the rate matrix  $J$  (Section 3.2), imply that the same trends found for the available information should be carried onto these metrics (improvement proportional to  $NE$ , bounding within the corresponding *S* and *US* “extremities”). Figure 4a illustrates the average predicted standard errors of estimates  $\hat{\pi}_1$  ( $\diamond$ ),  $\hat{\mu}_1$  ( $\nabla$ ) and  $\hat{\mu}_2$  ( $\square$ ) as well as their sum ( $\circ$ ) with increasing contextual negentropy  $NE$  for *US* (red), *CA* (blue) and *WCA* (green). Indeed, it is illustrated that standard errors of all parameters (see below for exceptions), as well as their sum, decrease with increasing negentropy for both algorithms that employ contextual assistance. They further converge towards a parameter-specific limit at  $NE = 1$  (which can be shown to be the standard errors of *S*). Conversely, they remain stable for *US* across all  $NE$  levels. Similarly, Figure 4b shows that in the same problem, the convergence rate  $r'$  of *CA* (blue) and *WCA* (green) is improved with increasing contextual negentropy to reach immediate convergence ( $r' = 1$ , single iteration just like the supervised algorithm) in the case of context fully revealing the missing labels ( $NE = 1$ ).

Another important result regards a notable difference between *CA* and *WCA*: As shown in Figure 4, standard errors and convergence rates of *CA* demonstrate a superiority over *WCA*, starting from a more favorable point at  $NE = 0$  and remaining better than the *WCA* equivalents for the largest part of  $NE$  spectrum, until, as discussed, both algorithms converge to the supervised equivalents at  $NE \rightarrow 1$ . *CA* is, hence, seemingly violating the previously extracted result of being strictly lower-bounded by *US* (by yielding a higher, more favourable lower bound).

7. It is intuitive and easy to show by means of the definitions of all the involved information matrices that  $I_c^{MAX} = I_c^{MIN} = I^S$ ,  $I_m^{MAX} = I_m^{US}$  and  $I_m^{MIN} = I_m^S = 0$ .

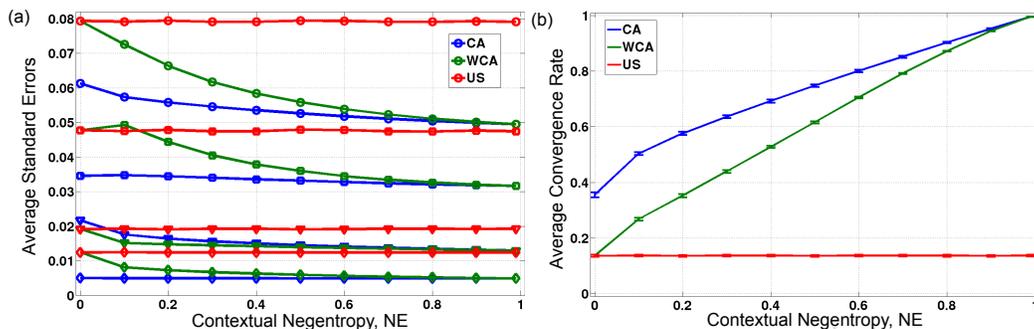


Figure 4: (a) Average predicted standard errors of  $\hat{\pi}_1$  ( $\diamond$ ),  $\hat{\mu}_1$  ( $\nabla$ ) and  $\hat{\mu}_2$  ( $\square$ ) and their sum ( $\circ$ ) with increasing contextual negentropy  $NE$  in 100 repetitions of a mixture of two univariate Gaussians estimation problem. Different algorithms color-coded in the legend. (b) Average predicted convergence rates and their standard deviations for the three algorithms (color-coded in the legend) in the same estimation problem.

The explanation of this effect can be found in the form of the complete-data log-likelihoods (Equations 21–25 in Appendix A) and the E-step formulations (Table 1). While the structure of the information matrices is identical between  $WCA$  and  $US$  for the same parameter estimation problem, the structure of the matrices for  $CA$  are *reduced* by removing the rows and columns corresponding to the mixing coefficients  $\pi_j, \forall j$  (Figure 3b, row and column for  $\pi_1$  are deleted). This comes from the definitions of  $I, I_c, I_m$  on the complete-data log-likelihoods of  $CA$ , where mixing coefficients are missing. Essentially, in  $CA$ , as betrayed by the corresponding graph (Figure 1a), the data priors (mixing coefficients) are not data-dependent, but fully determined by context/side-information through the probabilistic labels<sup>8</sup>. The overall fraction of missing information is thus reduced “de facto” for  $CA$  compared to  $US$  and  $WCA$  (for the same  $NE$ ), as a result of removing the missing information related to the mixing coefficients. It should be underlined that the total fraction of missing information (the spectral radius of the rate matrix  $J$ ) in some estimation problem will always reduce by fixing (i.e., removing from the estimation problem) one or more parameters. Consequently, the favourable lower bound of  $CA$  at  $NE = 0$  (“ignorant” context) still corresponds to that acquired by the  $US$  algorithm, however, in the “reduced” version of the estimation problem.

Another interesting result regards the fact that, exceptionally, the standard errors of parameter  $\mu_1$  (Figure 4a), unlike for the rest of the individual parameters and for their sum, do not demonstrate the aforementioned superiority of  $CA$  over  $WCA$ . Such exceptions can occur because the missing information is not necessarily distributed uniformly across the estimated parameters, or even identically among the different algorithms. However, their sum (the trace of the variance-covariance matrix) only depends on the overall fraction of missing information. The latter is shown to reduce with increasing  $NE$  and be smaller for the same  $NE$  level for  $CA$ , compared to  $WCA$ . Similarly, since the global and componentwise rates of convergence only depend on the total fraction of missing information<sup>9</sup>, too, the rate of convergence of  $CA$  is guaranteed to be higher than that of  $US$  and  $WCA$  for a given  $NE$ , as long as the mixing coefficients are included in the estimation problem.

8. Mixing coefficients  $\pi_j$  can still be computed in the  $CA$  case independently. The regular estimator  $\pi_j = \sum_{i=1}^N \mathbb{E}_{\hat{\theta}}\{z_{ij}\}/N$  can be shown to be biased even for the case of “correct” context. We will hereafter employ an alternative estimator as  $\pi_j = \frac{1}{N} \sum_i p'_{ij}$ , where  $p'_{ij} = 1$  if  $j = \operatorname{argmax}_k \{p_{ik}\}$  and 0 otherwise. The latter is unbiased and accounts for the same standard errors of mixing coefficients as for  $S$  (for “correct” context).

9. This is true except for rare cases, where different components/parameters can converge at different rates, see Meng and Rubin (1994).

### 4.3 Results on scenarios with artificial data

In Appendix C, it is formally shown that the conclusions of the previous section will generalize to arbitrary FMM cases. We experimentally verify the existence, magnitude, comparative statistical significance and particularities of the benefits expected by the analysis in Sections 4.1– 4.2, in several scenarios, employing the metrics and validation methodology described in Section 3.3. In parallel to these metrics, we also report the number of problems (out of 1000 problems solved for each scenario) that did not converge<sup>10</sup>.

#### 4.3.1 ESTIMATION SCENARIOS WITH “CORRECT” CONTEXT

The first set of artificial data simulations is meant to compare the performances of the proposed algorithms among each other and against  $S$  and  $US$  in terms of the variables of interest motivated above, on six scenarios that differ on the types and numbers of mixtures employed, the number of estimated parameters, the dimension of the input space and the utility of the FMM (classification versus regression).

Figure 5 reports the estimation precision as Euclidean distance  $D$  (see Section 3.3) between the actual FMM parameters and the ones estimated by each algorithm included in our comparative analysis. Note that the lower the value of this metric, the higher the estimation precision is. Firstly, as expected, statistically significant difference is showcased between  $S$  and  $US$  in all scenarios. As predicted,  $CA$  and  $WCA$  are able to increase estimation precision proportionally to the contextual information at hand ( $NE$  level), within the upper and lower bounds defined by  $S$  and  $US$ . The lower bound for  $CA$  is higher, corresponding again to the level acquired by the “reduced”  $US$  problem (with known mixing coefficients). It is extremely interesting that, although these results could only be exactly theoretically predicted for standard errors and convergence rates (Section 4.2), it is evident that the alleviation of missing information has the exact same impact on estimation precision, through the favourable distortion of the likelihood landscapes denoted in Section 4.1.

While the exact magnitude of these effects is scenario-dependent,  $CA$  and  $WCA$  are statistically significantly (red asterisks) superior to  $US$  at the lowest tested level of non-ignorant context tested ( $NE = 0.1$ ). For  $CA$ , this holds already for ignorant context. The only exception is the first scenario, where, since the mixing coefficient is not estimated but considered known, the relevant information is not missing either for the  $WCA$  algorithms. While, as noted, all context-aware algorithms converge towards the  $S$  levels for increasing  $NE$ , the difference becomes insignificant at best for  $NE = 0.6$  or above (yet, the magnitude of differences from  $S$  tends to be much smaller than the equivalent difference from  $US$ ). The superiority in terms of estimation precision of  $CA$  in comparison to  $WCA$  is not only true for ignorant context, but extends to all  $NE$  levels and scenarios (apart from, again, the first one) and tends to be significant (blue asterisks) for the first few  $NE$  levels.

$DCA$ , despite also improving proportionally to  $NE$  and converging towards the limit of supervised learning  $S$  for  $NE \rightarrow 1$ , significantly (for most of the  $NE$  spectrum) underperforms for low levels of contextual assistance, even compared to  $US$ . This result suggests that discarding *bottom-up* information completely is sub-optimal. This simple algorithm (which has the advantage of converging in a single iteration and demonstrating the same standard errors of  $S$ ) should be thus only employed when one enjoys contextual assistance of very high information content. The only exception is the mixture of regressions scenario (Figure 5f), where, while  $DCA$  still underperforms compared to the other context-aware algorithms, it outperforms  $US$  already for ignorant context. This effect is due to the fact that no minimum “separability” requirement between mixtures has been enforced in this scenario (see Appendix D), as it is fairly unintuitive to define a separability index between regression

10. Exceptions to the convergence of an EM algorithm to a local maximum occur when the spectral radius of  $J$  exceeds unity, implying the existence of a ridge in the incomplete-data log-likelihood. In this case, standard error estimates and convergence rates predicted through the evaluation of information matrices will not coincide with the measured values. In a problem converging to a local maximum, all eigenvalues of  $J$  lie in  $[0, 1]$  if  $I_c$  is positive semi-definite or  $[0, 1)$  if  $I_c$  is positive definite, see McLachlan and Krishnan (2008).

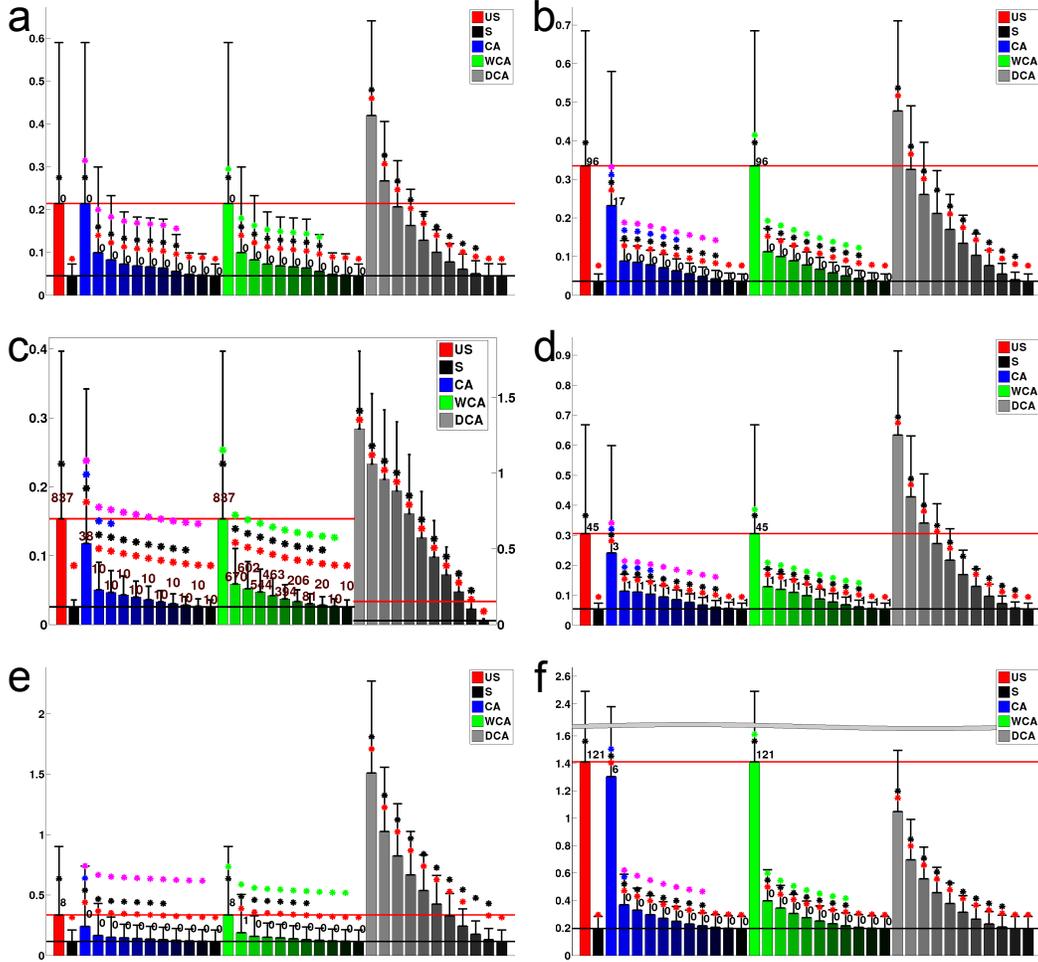


Figure 5: Averages and standard deviations of estimation precision  $D$  across 1000 problems for scenarios: (a) Mixture of two univariate normal distributions (only means  $\mu_1, \mu_2$  estimated). (b) Mixture of two univariate normal distributions. (c) Mixture of three univariate normal distributions. (d) Mixture of two multivariate (2D) normal distributions. (e) Mixture of two Maxwell–Boltzmann distributions. (f) Mixture of two univariate, first-order linear regressors. Different algorithms colour-coded in the legend. Context-aware algorithms for different  $NE \in [0 : 0.1 : 0.99]$  levels shown in shades of the respective algorithm’s color. On top of each bar, the number of problems that did not converge are shown. Coloured asterisks on top of each bar denote statistically significant difference across 1000 problems ( $\alpha = 0.01$ , Wilcoxon ranksum test), between the respective algorithm and:  $US$  in red,  $S$  in black, the equivalent  $NE$  level of  $WCA$  in blue (for  $CA$  algorithms only), the equivalent  $NE$  level of  $DCA$  in magenta (for  $CA$  algorithms only) and the equivalent  $NE$  level of  $DCA$  in green (for  $WCA$  algorithms only). The horizontal lines illustrate the level of  $US$  (red) and  $S$  (black). The right y-axis of (c) corresponds to  $DCA$ .

models. The lower the  $NE$ , the more the estimates of  $DCA$  for all mixtures can be shown to be biased towards the mixture-wise average of the true parameter values. These are incidentally close to the true values for low separability problems, what biases the estimation precision results of  $DCA$  in this scenario.

Context-awareness is also shown to substantially reduce the number of problems that could not converge with regular unsupervised learning  $US$ , a desirable effect which is again proportional to  $NE$ , and where  $CA$  once more outperforms  $WCA$ . Closer inspection of non-converged problems reveals that in the vast majority of cases, non-convergence is attributed to the irregularity of the log-likelihood functions rather than any insufficiency of the executed iterations. Hence, it can be said that context-awareness is able to turn irregular problems into regular ones, already at very low  $NE$  levels.

Figures 6-7 show the average (across converged problems) convergence rate  $r'$  and the equivalent average of “standard error average” (across parameters)  $ASE$ , respectively, for the same scenarios. Note that we avoid showing average  $r'$  for  $S$  and  $DCA$ , as it is always 1 (non-iterative algorithms), as well as the average  $ASE$  for  $DCA$ , as it is the same of  $S$ . These figures fully verify that the results of the example in Section 4.2 (benefits proportional to  $NE$  and bounded by  $US$  and  $S$ , superiority of  $CA$  over  $WCA$ ) indeed generalize to arbitrary FMM cases, and the proofs in Appendix C are experimentally substantiated. Significance is observed throughout the  $NE$  spectrum for  $r'$  and for at least the first few  $NE$  values for  $ASE$ .

While improvements in estimation precision are valuable per se in a variety of applications, especially wherever FMMs are employed for modeling and/or data generation purposes, it is true that, most commonly, such models are trained with the final goal of performing classification/regression tasks. It is reasonable to assume that improved estimation precision and reduced standard errors of estimation should have an impact on subsequent classification and regression. We therefore additionally study classification accuracy  $A$  (%) for all the above scenarios and  $MSE$  (for the mixture of regressions scenario), extracted by applying the trained models on a separate testing set, as described in Section 3.3.

Figure 8 illustrates that, on average, the impact of supervision (aka, missing information) on classification is important, since the supervised estimator,  $S$ , outperforms the unsupervised one,  $US$ , in all considered scenarios, the differences being statistically significant in all cases but the mixture of two Maxwell-Boltzmann distributions scenarios (Figure 8e). Still, in this case, an average difference of 5% in testing accuracy is derived across all problems solved.  $WCA$  for ignorant context performs identically to  $US$ , a direct consequence of the fact that these two algorithms are in fact identical and yield the same estimation precision. The derived superiority of  $CA$  for ignorant context in terms of estimation precision survives also in terms of  $A$  and  $MSE$  (with the already discussed exception of the first scenario, (a)). It is, however, statistically significantly better than  $US$  only for scenarios (b) and (f), while only insignificantly worse than  $S$  for scenario (e). The most interesting effect is that, although context-aware algorithms  $CA$ ,  $WCA$  still operate within the boundaries defined by  $S$  and  $US$ , much of the proportional relation to  $NE$ , evident in all metrics examined before, seems to vanish. Both algorithms perform very close to the supervised estimator already at  $NE = 0.1$ , with the exception of scenarios (a) and (e). As a result, apart from the latter scenarios, the superiority of  $CA$  over  $WCA$  is less evident (and not significant) in terms of classification/regression outcome. In any case, this can only be regarded as a positive effect, since large improvements, insignificantly different from supervised estimation  $S$  are brought forward (even for low  $NE$  levels) for both algorithms.

On the contrary, proportionality between  $NE$  and  $A/MSE$  is more evident for the  $DCA$  algorithm. Quite surprisingly, the compromised estimation precision of this algorithm for low  $NE$  does not translate into particularly poor classification/regression performance (except for ignorant context, where nearly chance-level classification is derived); in all cases,  $A$  and  $MSE$  are better than  $US$  even for low  $NE$  values. Still, algorithms  $CA$  and  $WCA$  outperform  $DCA$  significantly for low ranges of contextual assistance.

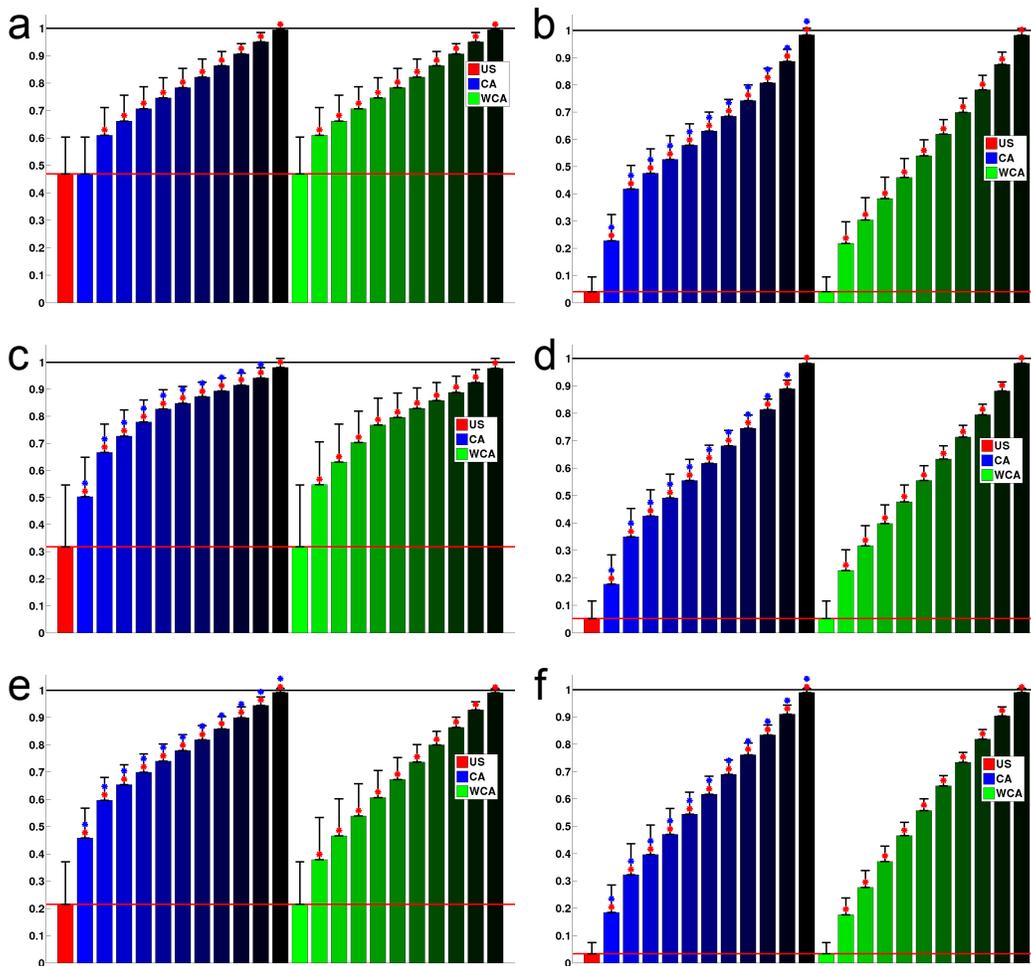


Figure 6: Average and standard deviations of convergence rate  $r'$  across all converged problems for all scenarios. All illustrations follow the same conventions of Figure 5.

#### 4.3.2 ESTIMATION SCENARIO WITH “MIXED” CONTEXTUAL INFORMATION

Having demonstrated the generalization of context-aware algorithms’ behavior in different FMM estimation scenarios, we limit our subsequent simulation studies (of various situations beyond “correct” context with uniform  $NE$ ) to a single scenario, involving mixtures of two univariate normal distributions. To begin with, the probabilistic label definitions in Table 1 already imply that, in real applications, it is highly unlikely that all labels will exhibit the same information content  $NE$ , or even that the latter will approach  $NE = 1$  (that would virtually correspond to an automatic label annotator). We therefore initially study the case of “mixed” and low context in the aforementioned scenario, where the information content of each probabilistic label is drawn randomly from a uniform distribution  $NE_i \in [0, 0.5]$ .

Figure 9 illustrates the average values of the four metrics of interest in this situation, across 1000 problems. Comparing Figure 9a to Figure 5b corresponding to the same scenario, the average estimation precision  $D$  for  $US$  and  $S$  is not altered, as expected, since these methods are context-independent. Similarly,  $CA$  remains slightly (but statistically significantly) superior to  $WCA$ , while both algorithms by far (and significantly) outperform  $DCA$ . It is worth to underline that, in this

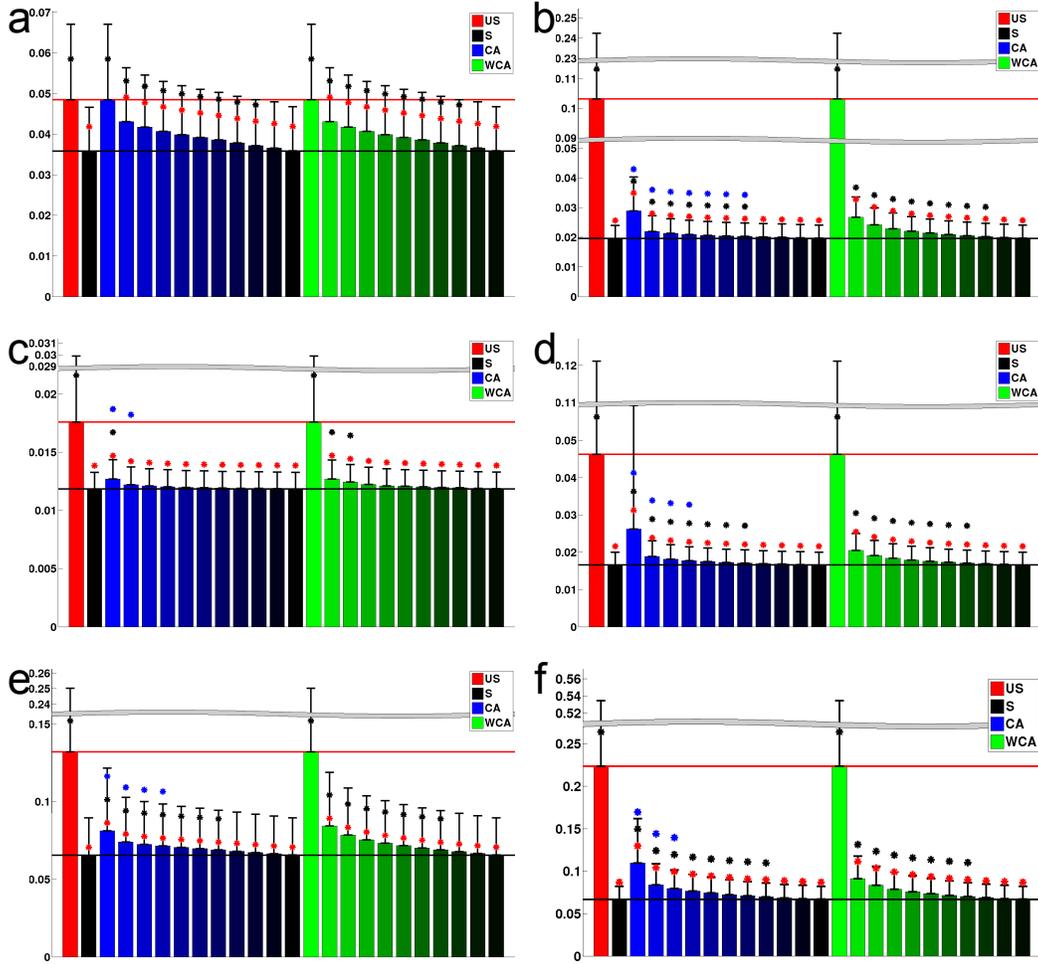


Figure 7: Average and standard deviations of average (across parameters) standard error  $ASE$  across all converged problems for all scenarios. All illustrations follow the same conventions of Figure 5.

case  $DCA$  remains at least superior to  $US$  (but not significantly). All algorithms remain significantly inferior to supervised learning  $S$ , yet, much closer to it than to  $US$ . It is also worth to note that the extracted estimation precision level for context-aware algorithms is very close to the average of  $NE$  levels 0-0.5 derived in the same scenario with “pure” context (Figure 5b). Essentially, it is shown that the latter type of simple simulations can be useful to predict benefits even in the case of “mixed” context, by performing a simulation using the average  $NE$  level one expects to retrieve in a specific application scenario. Additionally, both  $CA$  and  $WCA$  are shown to be able to cope with all 100 problems that did not converge in the  $US$  case.

Regarding convergence rate  $r'$  and average standard errors  $ASE$ , the same behaviour persists: the relative performances among algorithms are not altered by the situation of “mixed” context. Yet, the magnitude of effects tends to converge towards the one that would be obtained in a “pure” simulation, with a  $NE$  level corresponding to the average  $NE$  of labels in the “mixed” scenario. Finally, regarding classification accuracy, this “mixed” and low context situation already allows algorithms  $CA$  and  $WCA$  to operate equivalently to  $S$  and thus significantly better than  $US$  ( $DCA$  is slightly and

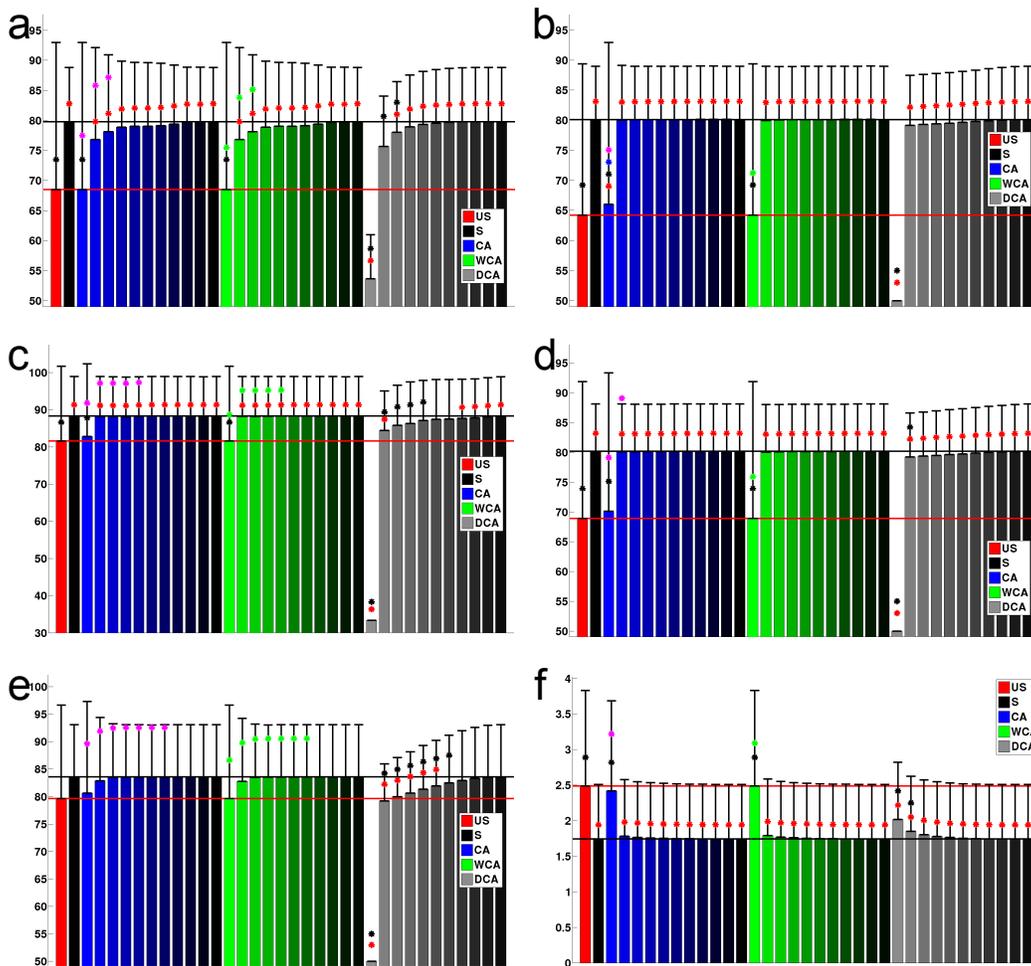


Figure 8: Average and standard deviations of testing set classification accuracy (%) A (a-e) and Mean Square Error (MSE) of regression (f) across 1000 problems for all scenarios. All illustrations follow the same conventions of Figure 5.

non-significantly inferior). This is an impressive outcome, revealing that large application-related benefits can be acquired without a requirement of very strong contextual assistance.

### 4.3.3 ESTIMATION SCENARIO WITH “WRONG” CONTEXT

Besides situations with “mixed” context, the definition of probabilistic labels in Table 1 provides additionally no guarantee that, in a real application, such labels will comply with the assumptions we have hereby termed “correct” context (Section 3.3) and applied so far. Hence, we evaluate next (in the same scenario of two mixtures of univariate normal distributions) nine different sets of 1000 problems each, modifying the percentage of “wrong” probabilistic labels (i.e., labels where the distribution of confidence across classes is not in agreement with the ground truth label) from 10% to 90% with a step of 10%. Figure 10 provides a concise color-coded illustration of the averages of

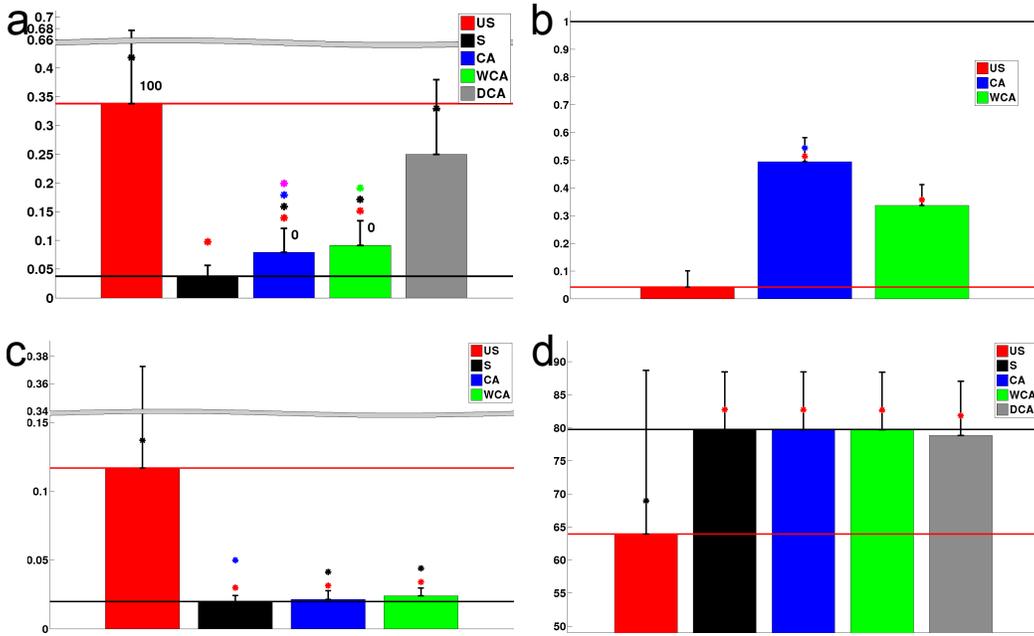


Figure 9: Average estimation precision  $D$  (a), convergence rate  $r'$  (b), average (across parameters) standard error  $ASE$  (c) and classification accuracy  $A\%$  (d) across 1000 problems of a single scenario with “mixed” context. All illustrations follow the same conventions of Figure 5.

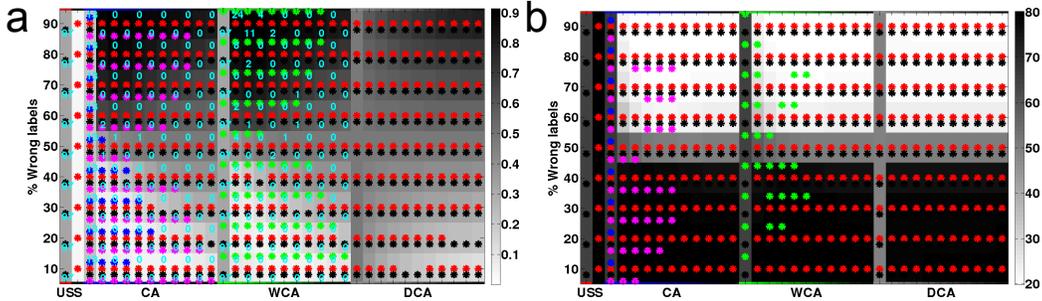


Figure 10: Average estimation precision  $D$  (a), and classification accuracy  $A\%$  (b) across 1000 problems of a “wrong” context scenario, solved for various percentages of wrong probabilistic labels (y-axis), and colour-coded as shown in the respective colourbars. Different algorithms and  $NE$  levels on the x-axis. All other illustrations follow the same conventions of the previous figures.

estimation precision  $D$  and classification accuracy  $A$  for all algorithms (columns) and intensity of “wrong” context (rows)<sup>11</sup>.

Figure 10a shows that, for percentages of erroneous contextual assistance at 60% and above, all context-aware algorithms perform significantly worse than  $US$  (naturally also  $S$ ) concerning estima-

11. To keep these illustrations concise, the information of standard deviation of each metric across the 1000 problems is not depicted in these maps.

tion precision. That, of course, excludes the ignorant context algorithms for  $CA$  (improved compared to  $US$ ) and  $WCA$  (identical to  $US$ ), since for ignorant context (uniform probabilistic labels), there can be no distinction between “correct” and “wrong” context. This result (also leading to “inverted” classification accuracy for all context-aware algorithms in Figure 10b) verifies experimentally that it is nonsensical to employ side-information that is by majority inaccurate.

For erroneous labels at 50% and below, context-aware algorithms yield significant improvements over  $US$  throughout the  $NE$  spectrum. However,  $CA$  and  $WCA$  exhibit deteriorating precision for increasing  $NE > 0$  (and the same percentage of “wrong” labels). Albeit initially counter-intuitive, this effect has a reasonable explanation. The more “confident” (high  $NE$ ) the inaccurate probabilistic labels are, the more harmful they will be. Consequently, for 50% “wrong” labels, context-aware algorithms still underperform against  $US$  throughout the  $NE$  spectrum. However, for percentages of “wrong” labels at 40%, there already exists some  $NE$  value below (and not above!) which, context-aware algorithms significantly outperform  $US$ . These information content levels are  $NE = 0.5$  at 40% “wrong” context for both  $CA$  and  $WCA$ , and  $NE = 0$  for lower percentages of “wrong” context (with, as said, decreasing magnitude of improvement as  $NE$  increases). Both  $CA$  and  $WCA$  are thus shown to achieve improved estimation precision over completely unsupervised learning already at 40% of “wrong” context for low  $NE$  and to be completely superior for all  $NE$  levels at “wrong” context below 30%. Despite significant superiority to  $US$ , these improvements are naturally of lesser magnitude compared to the “correct” context scenario in Figure 5b.  $DCA$  is the only algorithm that maintains the same behavior of “correct” context, where improvements are still proportional (and not inverse proportional) to  $NE$ . Yet, these improvements only become superior to  $US$  for high contextual information  $NE$  and small percentages of “wrong” context, while also being significantly inferior to the equivalent improvements of both  $CA$  and  $WCA$ .

Convergence rates  $r'$  and average standard errors  $ASE$  (illustrations not shown for brevity) are not particularly affected by “wrong” context, demonstrating similar effects to the “correct” context scenario. The only interesting finding regards the fact that, the magnitude of improvements on convergence rate, depends on the intensity of “wrong” context, for both  $CA$  and  $WCA$ . The improvements are found to be greater for percentages of “wrong” probabilistic labels approaching 50%. A theoretical justification of the effects of “wrong” context on these metrics through the MIP cannot be easily established at this point. Yet, this is an interesting topic for future work.

Classification accuracy  $A$  (%) in Figure 10b is shown to follow exactly the aforementioned effects in estimation precision, as well as the same attenuation of the dependence to  $NE$  already concluded for “correct” context in Section 4.3.1 and Figure 8b. As a result, for problems where the majority of labels are wrong,  $A$  is below chance level and, in fact, “opposite” to the performance of  $S$ . This is obviously the effect of a class-inversion that occurs in the 2-class problem of our scenario, when the majority of labels point towards the opposite class. At 50% “wrong” context,  $A$  is shown to be on average around chance level. Symmetrically, even for large (but, still, minority) percentage of “wrong” probabilistic labels, the accuracy is very close to that achieved with supervised learning  $S$  for all context-aware algorithms, even  $DCA$ .  $CA$  cannot outperform  $WCA$  apart from the known ignorant context case, yet, both these algorithms significantly outperform  $DCA$  for low  $NE$ .

Concluding, the existence of a minority amount of “wrong” probabilistic labels among the ones that one can retrieve in a real-world problem, is by no means detrimental to the application of context-aware algorithms. That holds even when the percentage of “wrong” labels approaches 50%. In this case, one should be aware that low information content of labels around  $NE = 0.6$  actually yields superior results to “strong” contextual assistance. This is an overall advantageous effect, given that in real applications, one is more likely to be able to retrieve context of low information content. Still, improvements in the presence of “wrong” context are lesser than those acquired in the ideal, “correct” context situation.

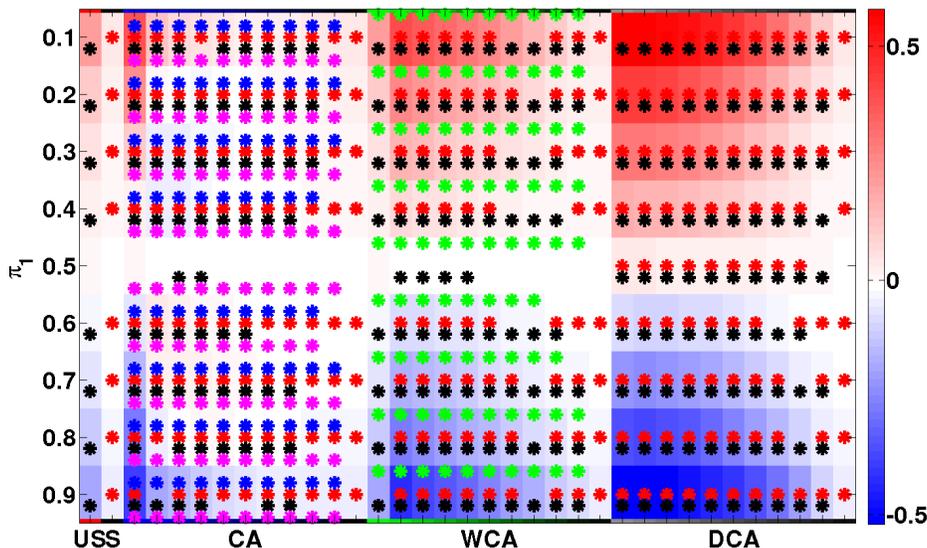


Figure 11: Average estimation class-bias  $B$  across 1000 problems of a “biased” scenario, solved for various values of mixing coefficient  $\pi_1$  (y-axis). All other illustrations follow the same convention of the previous figure.

#### 4.3.4 ESTIMATION SCENARIO WITH UNBALANCED NUMBER OF SAMPLES PER CLASS

Returning to situations with “correct” context, the simulation results presented in Section 4.3.1 only concerned “balanced” problems, i.e., problems where the same number of training samples is provided for all classes. In unbalanced problems that can frequently occur in practice, estimation precision is known to be superior for the dominating class(es), an effect attributed to the fact that the missing information for this class(es) is comparatively reduced. Since, context-awareness is shown to have a direct impact on the amount of missing information, it is reasonable to imagine that it could have a beneficial effect in coping with such class-biasing phenomena. We evaluate the impact of the proposed context-aware algorithms in such situations (and in the same scenario of mixtures of two univariate normal distributions) running 9 sets of 1000 problems, manipulating for each set the actual mixing coefficient  $\pi_1$  within  $[0.1, 0.9]$  (and, thus, also the percentage of training samples belonging to the first class) with a step of 0.1.

As expected, in terms of dependence on  $NE$  and type of algorithm, the average values of all metrics studied here ( $BA$  instead of  $A$  has been used for classification performance in this situation, as justified in Section 3.3), across 1000 problems and for each  $\pi_1$  value imposed, demonstrate identical trends to the equivalent “correct” context scenario studied in Section 4.3.1 (Figures 5b-8b). The corresponding illustrations are omitted in the interest of the manuscript’s economy. Comparisons across different  $\pi_1$  values (“bias” levels), however, show that class-bias does indeed have a detrimental effect on all these metrics for the  $US$  algorithm (considerably also on the number of converging problems), which are more obvious regarding estimation precision and convergence rate. Supervised learning  $S$  is also naturally affected, but to a lesser extent. Context-aware algorithms can be shown to significantly alleviate the negative impact of class-bias for all metrics tested. The benefits are proportional to the  $NE$  level of contextual assistance and larger in magnitude the higher class-bias levels are.

In order to make the impact of context-awareness on class-bias alleviation clearer, Figure 11a illustrates an average estimation bias metric  $B$ , as the difference in estimation precision (quantified, again, with metric  $D$ ) between the parameters of each mixture individually ( $B = D_1 - D_2$ ). Shades

of blue and red denote superior estimation precision in favor of the first or second class, respectively, as color-coded in the corresponding colorbar. The first column verifies that in regular unsupervised learning, *US*, the unbalanced number of samples will indeed favor the estimation precision of the dominating class’ parameters, an effect augmenting in magnitude proportionally to the absolute difference (bias level)  $\|\pi_1 - 0.5\|$ . All context-aware algorithms are shown to be able to remove the biasing effect for all bias levels tested, up to the upper bound defined by supervised learning *S*. *CA* achieves remarkable results already for low *NE* levels, being far (and statistically significantly) superior to *WCA*, while, once again, both these algorithms significantly outperform *DCA*. Concluding, context-aware algorithms are proved to yield significant benefits in coping with estimation bias in unbalanced problems, a point where *CA* is shown to be again significantly superior among the proposed algorithms.

#### 4.4 Online context-aware learning in brain-computer interface

The applicability and effectiveness of context-aware learning are demonstrated in an online-learning problem from the field of brain-computer interaction (BCI). More specifically, we employ the binary “BrainTree” speller described in Perdakis et al. (2014), which allows a user to type messages by means of two control commands. A 2-class, motor imagery (MI) BCI translates processed brain activity patterns monitored through electroencephalographic (EEG) signal into one of the two required application control commands, as in Leeb et al. (2013). Such an application is ideal for demonstrating the benefits of context-aware learning; on one hand, BCI is known to suffer from non-stationarity of the extracted brain patterns, what degrades previously trained classifiers and calls for online classifier learning (Millán, 2004). Yet, the latter has to be carried out in an unsupervised manner since data labels cannot be retrieved during online BCI operation. Consequently, adaptive classifier training in BCI is bound to suffer the known shortcomings of unsupervised learning. On the other hand, the existence of a smart brain-actuated device in the control loop provides a natural candidate for the extraction of contextual assistance. It is thus expected that context-aware learning, as proposed in this work, could allow uninterrupted BCI spelling avoiding both the deficiencies of unsupervised learning and the lengthy supervised retraining sessions.

Figure 12a (top) illustrates the speller’s graphical user interface (GUI), where characters are arranged alphabetically. The vertical red cursor (“caret”) denotes the current position in this character bar, while the orange “bubble” surrounds currently available characters. Underneath, the user observes a conventional MI BCI feedback, consisting of a green cursor extending left/right within a feedback bar. Regarding the control paradigm, as soon as the user has identified the position of the desired character relative to the “caret” (left/right), he/she employs the respective MI task (e.g., imagination of right/left hand movement) to extend the feedback cursor towards the desired side. The feedback cursor extends left/right according to the BCI’s classification outcome on two brain patterns associated to the two MI tasks. As soon as the cursor has reached a threshold (blue edges of the feedback bar), the “caret” moves towards this side and closer to the desired character. The procedure is repeated until the desired character is the only one left within the orange bubble, in which case the next movement will append it to the typed text area. A new typing round is then initiated, where the orange “bubble” will surround again all available characters. This simple GUI simplifies the speller’s underlying structure, where characters are the leaf nodes of a binary tree (example on a reduced in Figure 12a, bottom). Thus, the caret’s position is simply the current internal node and the orange “bubble” surrounds the leaf node characters belonging to the current node’s two subtrees. Left/right BCI commands move the current node to the left/right child of the current node. A new tree is generated after a character is typed, setting the “caret” to the root. Effectively, in each typing round, each character is associated with a binary “codeword” of left/right transitions.

This underlying structure provides a straightforward mechanism for retrieving contextual assistance through the speller and applying *CA* learning<sup>12</sup>. By modeling the desired character as a contextual random variable  $c \in [a, b, \dots, z, \text{space}, \text{backspace}]$ , according to the probabilistic label definitions (Table 1), one only needs to know the priors  $p(c)$  and conditionals  $p(z|c)$  (where  $z \in [0, 1]$ , the MI class the user is employing). A trained Prediction by Partial Matching (PPM) language model provides the priors  $p(c)$  for each typing round, based on the currently written prefix. Conditionals  $p(z|c)$  are also easily extracted given the structure of the tree and knowledge on the current node position, information which is always readily available. More specifically,  $p(z = j|c = k) = 1$  holds if the character  $k$  is a member of the subtree  $j \in [\text{left}, \text{right}]$  and  $p(z = j|c = k) = 0$ , otherwise.

A custom type of binary tree able to provide implicit probabilistic labels of high information content, *NE*, is employed. Each node’s subtrees are arranged to obey as much as possible a 0.9/0.1 or 0.1/0.9 split of total character probability (the “heavy” subtree is reversed at each level of the tree to avoid a “class–correlated” context situation), while still maintaining alphabetic ordering. This results in a “mixed” context scenario, since the aforementioned split is not always possible given any position in the tree and the current  $p(c)$ . A small percentage of “wrong” context also exists, since it is not guaranteed that the desired character is indeed a member of the “heavy”, most probable subtree.

We devise a buffer approach for continuous, context–aware, online–learning of a BCI classifier, modeled as a mixture of two multivariate, 6–dimensional normal distributions with common covariance matrix. Effectively, that is an LDA classifier. Six features capturing a subject’s spatially distributed sensorimotor rhythms are extracted in a sliding window, twice per second (2 Hz). EM–learning takes place in a buffer of the latest two minutes of data (240 feature vectors/samples). Consecutive buffers are shifted by only 1 sample, thus a new, slightly updated classifier is used to classify each incoming sample, reflecting the evolving brain patterns of the recent 2 minutes. The spatio-spectral features extracted for each MI task are log-transformed and known to be approximately normally distributed (Kolmogorov-Smirnov test, 95% confidence interval).

We conduct spelling simulations using EEG MI data of 12 subjects recorded in the lab with a conventional, 2-class BCI protocol described by Leeb et al. (2013). For simulated spelling with each evaluated algorithm, each subject’s data are “played–back” in the order recorded; specifically, when a subject would need to go right/left for reaching the desired character, the earliest samples of the first/second (respectively) MI task not used thus far in the subject’s dataset are forwarded to the adaptive BCI. For all subjects, a common subject–unspecific classifier is used as the initial point of adaptation. The same spelling task of typing the words “nothing” and “portion” is repeated for the supervised algorithm *S* (true data labels used), the *CA* case and the *CA* with ignorant context (noted *CAE*, constant  $NE = 0$ ). Automatic correction of erroneous commands is imposed, simulating the hybrid correction mechanism that actual users could employ in the non-adaptive version of the speller (Perdikis et al., 2014). Upon an erroneous command, the “heavy” side of the two subtrees is reversed. The number of samples per class in each buffer is variable, depending on the variable numbers of left/right commands within the last two minutes and of samples necessitated for each command. Yet, since the average number of samples per command is much smaller than the buffer size (240), and by reversing the “heavy” subtree’s side upon errors, no particular class–bias exists.

Classification accuracy improvement over *CAE* is expected for the *CA* case, mainly as a result of its previously demonstrated superior estimation ability for increasing *NE*. Additionally, due to the fact that limited number of EM iterations are allowed (as many as can be executed within 100 msec) to cope with the real–time demands of the application; the higher convergence rate of *CA* for increasing *NE* is thus also expected to have a positive impact on classification accuracy. *US* has been found to yield nearly chance–level accuracies for all subjects, as a result of considerably compromised estimation precisions. This is reasonable since a large number of parameters needs

12. *CA* learning with unobserved context is applied, since the speller is unaware of the user’s desired character. *WCA* is here inapplicable, first, because it requires observed context and, second, because in BrainTree,  $z$  (MI task/desired direction of transition) depends on  $c$  (desired character), as in *CA* (Figure 1a).

to be learned from only 240 samples in each consecutive buffer and only a few allowed iterations. *CAE* is hereby tested as a slightly improved “surrogate” of *US*. This application resembles that of Kindermans et al. (2012a,b). We calculate “balanced”, running 2-class classification accuracy in a window of the latest minute (120 samples) of simulated BCI spelling, with a shift of 30 seconds (60 samples), a sort of prequential evaluation akin to online learning. Class unbalance is not particularly intense, for the same reasons mentioned regarding the adaptation buffer. We nevertheless report “balanced” accuracy *BA* to take into account any class-bias effects.

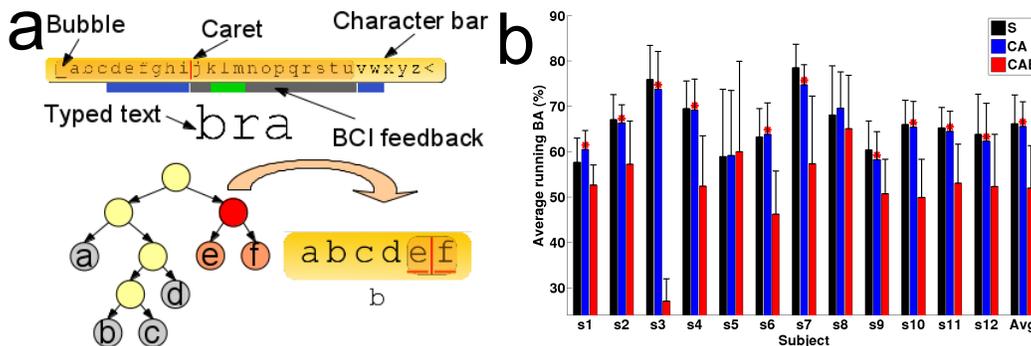


Figure 12: (a) Graphical user interface of the BrainTree speller (top) and its underlying binary tree structure. (b) Mean and standard deviation of balanced running classification accuracy during the spelling tasks for each of the 12 subjects with supervised estimation *S*, as well as algorithms *CA* and *CAE*. Red asterisks on top of *CA* bars denote statistical significant difference with *CAE* (Wilcoxon ranksum test,  $\alpha = 0.01$ ). No statistical significant differences between *CA* and *S* are found. The last triplet of bars illustrates the averages across subjects.

Figure 12b shows the averages and standard deviations of running *BA* after the first minute of spelling until the end of the spelling task. The large standard deviations even for *S* reflect the fact that brain patterns are intensively non-stationary for all subjects. Yet, *CA* learning is shown to quickly recover from the naive initial classifier, yielding similar *BA* to the supervised case, an overall astonishing outcome. That holds for each subject individually, as well as for the average across subjects. Classification results of *CA* are considerably better than those for the ignorant-context *CAE* algorithm, which is close to chance level for most subjects. The case of subject s5, where all algorithms perform similarly, was found to be due to intense instabilities of the generated brain patterns. Furthermore, *CAE* for subject s3 demonstrates a class-inversion effect, which is attributed to the fact that, although this subject is able to generate discriminant and fairly stable brain patterns, algorithm *CAE* cannot escape the local maxima near the initialization point throughout the spelling session. On the contrary, in the case of subject s8, initialization is coincidentally favourable, thus *CAE* is only slightly and non-significantly inferior.

Overall, the above example showcases the possibility to intuitively apply context-aware parameter estimation in real-world applications, involving tough classification and, possibly, online-learning problems. The demonstrated combined benefits in EM-MLE parameter estimation (faster convergence rate, suppression of unfavorable local maxima, less sensitivity to initialization, operation close to the supervised case) yield a significantly improved classification outcome without explicit supervision or otherwise manual labour to collect data labels. Improvement is achieved despite the fact that a non-ideal case of “mixed” and “wrong” context case is studied. In the particular application in question, context-aware learning will allow users to communicate with a BCI-speller “on-demand”, without the need of supervised retraining sessions (while achieving almost the same benefits of those)

that severely limits the deployment of BCI in the assistive technology market and every-day life of disabled individuals.

## 5. Summary and discussion

This work has presented a comparative analysis and in-depth study of the properties of maximum-likelihood estimation algorithms for finite mixture models. These algorithms reflect an “atypical” semi-supervised setting, ignoring the true data labels but exploiting a weaker type of supervision in the form of implicitly extracted “soft” labels. The latter are derived through probabilistic context modeling, where context is embedded into the model. Additional side-information results from knowledge on the prior and conditional distributions relating context to the latent data. It is of utmost importance to stress on the fact that the proposed framework, unlike most of the relevant works discussed in Section 2, requires absolutely no explicit labeling of any kind. It is hence an improved alternative of unsupervised learning, aside its obvious relation to methods that exploit side-information.

The first contribution of this article regards the derivations of the proposed algorithms (Section 3.1 and Appendix A), which, first, establish their soundness and non-heuristic nature and, second, justify their property of “context-awareness”. The former offers a robust (and devoid of any manual labeling need) alternative interpretation of such methods—recall that *WCA* has also been derived in other settings by Côme et al. (2009); Bouveyron and Girard (2009). The latter allows the definition of probabilistic labels to obtain a physical meaning in particular applications. These derivations thus dictate which of the proposed algorithms should be employed, given the particular structure of a learning problem (i.e., the relationship between context and latent class labels).

Another contribution entails revealing the two basic principles through which context-aware EM-MLE is able to yield significant benefits across various metrics. The first principle, concerning the estimation precision, regards the “distortion” of a context-assisted log-likelihood objective function (in comparison to the regular unsupervised one) in a way that a local maximum close to the supervised estimate is enhanced and the function’s values for the remaining parameter space (including other local maxima) are suppressed. The chances of a context-aware algorithm to converge closer to the supervised MLE are hence increased, while the importance of initialization diminishes. The second principle, directly influencing standard errors and convergence rates (indirectly, also estimation precision), regards the partial elimination of missing label information through context as a result of the applicability of the missing information principle. Through the above principles, we have established experimentally and, wherever possible, also formally, two important points. First, that the positive effects on all examined metrics are proportional to the information content of the extracted instance-wise probabilistic labels. Second, that the proposed algorithms will perform between the boundaries defined by the unsupervised and supervised equivalents of a given problem.

Demonstrating the applicability of the MIP makes our work the first one, among many investigating the effects of side-information in learning, to explain the experimentally derived benefits from an information-theoretic perspective. Future work could enhance this line of research in various directions, e.g., theoretically explaining the effects of imprecise additional information. It should be noted that throughout this work we make the regular assumption that missing data (i.e., ground truth labels) are “missing at random” (MAR), and thus, not correlated to the ground truth labels or any other variable. This assumption is very common and EM algorithms are known to perform reasonably well even in violation of it. Testing the effects of our algorithms in violation of the MAR assumption presents another interesting avenue for future work.

An additional contribution of our work is the exhaustive experimental verification of the theoretically expected benefits of context-aware algorithms by the results of simulation studies in different estimation scenarios. It has been demonstrated that contextually assisted learning demonstrates improved estimation precision, standard errors, convergence rates and classification/regression performances for different numbers and types of mixtures and estimated parameters. Additionally, our

results show that context-aware algorithms are able to alleviate class-bias in unbalanced problems. We have showcased that all benefits are still evident and significant in problems with variable and low contextual negentropy, or “wrong” context (when imprecise probabilistic labels are not dominating). In all these situations, we have identified potential limitations of the proposed algorithms. These conclusions are of great importance, as such problems are more likely to arise in practical applications. Finally, our results support the fact that context-awareness is able to turn irregular, non-converging problems into regular ones, which can also be expected by application of the first operation principle.

A fourth contribution regards the conclusions of the comparative analysis of the context-aware methods considered here. First of all, we have shown that method *DCA* achieves very biased estimates for low contextual negentropy, while, on the other hand, converging immediately in a single iteration. Disregarding bottom-up information derived from the observed samples to form the expectations over the latent variables is thus shown to yield inferior estimation properties. The superiority of *CA* over *WCA* as a result of removing missing information related to the mixing coefficients has also been demonstrated, affecting all metrics considered here. *CA* is also shown to be more effective in coping with class-bias.

Furthermore, the application of the *CA* algorithm in a tough, real-world, online-learning problem proves both the broad applicability of context-aware learning as well as the fact that the aforementioned improvements, combined, yield improved MLE and as a result superior online classification performance. Concerning general applicability, the tools discussed in this work (information matrices) provide the potential user with the means to estimate the benefits of context-aware algorithms and then compare them to the cost of retrieval to proceed with optimal application design. This provides an additional motivation for studying the benefits of context-aware learning against increasing contextual information content.

The main limitation of the algorithms proposed here is their non-universal applicability. Indeed, it is not guaranteed that for any application exploitable context exists, or that the cost of automatically retrieving contextual assistance (at least, at the required degree of information content) will be less than that of explicitly labeling data. However, it is evident by the provided real-world example that rich context can be easily and cheaply acquired in a broad application spectrum. The intuitiveness of probabilistic context modeling proposed here assists towards this direction.

Another criticism on this work could refer to a potential limitation of its applicability on FMMs. First, it should be underlined that proving the generalization of our claims to all types of FMMs already accounts for extensive applicability of the presented algorithms; these models are in themselves a very general tool, by virtue of the possibility to replace the mixture types, number of mixtures, etc., to the ones suited to a particular problem. The demonstrated applicability of our algorithms to the mixture-of-regressions scenario supports this claim. Furthermore, the principles of probabilistic context modeling embedded to a graphical model and of derivation by standard EM methodology, imply that similar benefits can be derived in more complex Bayesian networks. The only prerequisite is ensuring that the contextual variables with which a probabilistic graph is augmented, allow additional information flow towards latent nodes. The proposed algorithms are thus only an example of a broader class of algorithms that can be established.

The place of the proposed algorithms in the literature of learning with side-information should be further discussed. A major distinction to the majority of methods mentioned in Section 2 is that our framework does not need any kind of manual labeling, not even one that is “noisy”, uncertain, derived through crowd-sourcing, or otherwise. We have further focused on four frameworks which are the most generic and shown to be equivalent to each other (Ganchev et al., 2010). In Appendix B, we have drawn the analogy of our algorithms to Posterior Regularization (PR) and acknowledged that PR and similar frameworks are more generic and able to handle a larger variety of learning problems, by avoiding the modeling restrictions imposed here. Yet, we have argued that our own algorithms are, first, generic enough in themselves, more parsimonious, in that they require much simpler derivations and formalizations, and, evidently, much more intuitive and better suited for a

large variety of problems where side-information is easily represented probabilistically and can be embedded into the model. These facts allowed us to naturally derive two different algorithms (*CA*, *WCA*) and straightforwardly apply the MIP.

Concluding, our work has further established that the concept of context-awareness can play a key role in model learning, beyond the scope of inference where it has been most commonly employed so far. Future work could entail investigation of the effects context-aware learning might have in Bayesian estimation, as well as how our conclusions generalize for models other than FMMs.

## Acknowledgements

This work was supported by the European ICT Programme Project FP7-224631 and the Hasler Foundation, Switzerland. This paper only reflects the authors views and funding agencies are not liable for any use that may be made of the information contained herein.

## Appendix A. Derivations of *CA* and *WCA* parameter estimation

We derive the expressions for the incomplete- and complete-data likelihoods, the E-step and the probabilistic labels  $\mathbf{p}_i$  (Table 1) for methods *CA* and *WCA* through the directed graphs of Figure 1.

The joint distributions of the extended FMM graphical models of Figure 1 are:

$$CA \rightarrow p(\mathbf{x}_i, z_i, c_i) = p(c_i)p(z_i|c_i)p(\mathbf{x}_i|z_i) \quad (6)$$

$$WCA \rightarrow p(\mathbf{x}_i, z_i, c_i) = p(z_i)p(c_i|z_i)p(\mathbf{x}_i|z_i) \quad (7)$$

for *CA* and *WCA*, respectively. The joint distributions  $p(\mathbf{x}_i, c_i)$  can be derived by marginalizing out  $z_i$ :

$$CA \rightarrow p(\mathbf{x}_i, c_i) \stackrel{(6)}{=} p(c_i) \sum_{z_i} p(z_i|c_i)p(\mathbf{x}_i|z_i) \quad (8)$$

$$WCA \rightarrow p(\mathbf{x}_i, c_i) \stackrel{(7)}{=} \sum_{z_i} p(z_i)p(c_i|z_i)p(\mathbf{x}_i|z_i) \quad (9)$$

and the distribution of  $\mathbf{x}_i$  conditioned on  $c_i$  will be  $p(\mathbf{x}_i|c_i) = p(\mathbf{x}_i, c_i)/p(c_i)$ , thus:

$$CA \rightarrow p(\mathbf{x}_i|c_i) \stackrel{(8)}{=} \sum_{z_i} p(z_i|c_i)p(\mathbf{x}_i|z_i) \quad (10)$$

$$WCA \rightarrow p(\mathbf{x}_i|c_i) \stackrel{(9)}{=} \sum_{z_i} p(z_i) \frac{p(c_i|z_i)}{p(c_i)} p(\mathbf{x}_i|z_i) \quad (11)$$

while, for *CA* with unobserved  $c_i$ ,  $p(\mathbf{x}_i) = \sum_{c_i} p(\mathbf{x}_i, c_i)$ , so from Equation 8:

$$CA \rightarrow p(\mathbf{x}_i) \stackrel{(8)}{=} \sum_{c_i} p(c_i) \sum_{z_i} p(z_i|c_i)p(\mathbf{x}_i|z_i) = \sum_{z_i} p(\mathbf{x}_i|z_i) \sum_{c_i} p(c_i)p(z_i|c_i) \quad (12)$$

By conditioning on  $c_i$  (when observed) or marginalizing it out (latent) in Equations 6 and 7, we obtain:

$$CA, \text{ observed } c_i \rightarrow p(\mathbf{x}_i, z_i|c_i) \stackrel{(6)}{=} p(z_i|c_i)p(\mathbf{x}_i|z_i) \quad (13)$$

$$WCA, \text{ observed } c_i \rightarrow p(\mathbf{x}_i, z_i|c_i) \stackrel{(7)}{=} p(z_i) \frac{p(c_i|z_i)}{p(c_i)} p(\mathbf{x}_i|z_i) \quad (14)$$

$$CA, \text{ latent } c_i \rightarrow p(\mathbf{x}_i, z_i) \stackrel{(6)}{=} \left[ \sum_{c_i} p(z_i|c_i)p(c_i) \right] p(\mathbf{x}_i|z_i) \quad (15)$$

At this point, we proceed with the following definitions of probabilistic labels  $\mathbf{p}_i = \{p_{i1}, \dots, p_{ij}, \dots, p_{iM}\}$ :

$$CA, \text{ latent } c_i \rightarrow \mathbf{p}_i = \sum_{c_i} p(c_i) p(z_i|c_i) \quad (16)$$

$$CA, \text{ observed } c_i \rightarrow \mathbf{p}_i = p(z_i|c_i) \quad (17)$$

$$WCA, \text{ observed } c_i \rightarrow \mathbf{p}_i = \frac{p(c_i|z_i)}{p(c_i)} \quad (18)$$

Therefore, the definition of probabilistic labels through context requires in the *CA* case known distributions  $p(z_i|c_i)$  if  $c_i$  is observed and additionally known prior  $p(c_i)$  if  $c_i$  is latent. In the *WCA* case  $c_i$  should be observed (otherwise  $p(\mathbf{x}_i)$  reduces to the normal *US* case and no additional information exists) and known priors  $p(c_i)$  and conditionals  $p(c_i|z_i)$ . Note that without loss of generality  $c_i$  are assumed to be discrete random variables. In the continuous case summation over  $c_i$  should be replaced by integration.

The incomplete-data likelihood can be derived as  $\prod_{i=1}^N p(\mathbf{x}_i)$  or  $\prod_{i=1}^N p(\mathbf{x}_i|c_i)$  for latent and observed  $c_i$ , respectively. Replacing definitions 16-18 into Equations 10, 11 and 12, and  $p(\mathbf{x}_i|z_i) = f_j(\mathbf{x}_i|\theta_j)$ , one obtains the incomplete-data likelihoods:

$$CA \rightarrow L(\theta) = \prod_{i=1}^N \sum_{j=1}^M p_{ij} f_j(\mathbf{x}_i; \theta_j) \quad (19)$$

$$WCA \rightarrow L(\theta) = \prod_{i=1}^N \sum_{j=1}^M p_{ij} \pi_j f_j(\mathbf{x}_i; \theta_j) \quad (20)$$

and the logarithm of these expressions leads to the incomplete-data log-likelihoods of Table 1 for both methods.

We now proceed with the extraction of the complete-data log-likelihoods for the three cases under consideration. For *CA* with observed  $c_i$ , from Equations 17 and 19 and introducing indicator variables  $z_{ij}$  as in Equation 2:

$$\begin{aligned} CA, \text{ observed } c_i \rightarrow L_c &= \prod_{i=1}^N \prod_{j=1}^M [p(z_i|c_i) p(\mathbf{x}_i|z_i)]^{z_{ij}} \Rightarrow \\ \log L_c &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p(z_i|c_i) p(\mathbf{x}_i|z_i)) = \\ &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p(z_i|c_i)) + \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p(\mathbf{x}_i|z_i)) = \\ &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log p_{ij} + \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(f_j(\mathbf{x}_i; \theta_j)) \end{aligned} \quad (21)$$

and taking the expectation  $\mathbb{E}_{\hat{\theta}}\{\log L_c\}$  yields the expected complete-data log-likelihood for *CA*:

$$Q(\theta, \hat{\theta}^k) = \sum_i \sum_j \mathbb{E}_{\hat{\theta}^k}\{z_{ij}\} \log p_{ij} + \sum_i \sum_j \mathbb{E}_{\hat{\theta}^k}\{z_{ij}\} \log(f_j(\mathbf{x}_i; \theta_j)) \quad (22)$$

Similarly, for the WCA case, from Equations 18 and 20 using the same trick of indicator variables:

$$\begin{aligned}
 WCA, \text{ observed } c_i \rightarrow L_c &= \prod_{i=1}^N \prod_{j=1}^M \left[ \frac{p(c_i|z_i)}{p(c_i)} p(z_i) p(\mathbf{x}_i|z_i) \right]^{z_{ij}} \Rightarrow \\
 \log L_c &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log \left( \frac{p(c_i|z_i)}{p(c_i)} p(z_i) p(\mathbf{x}_i|z_i) \right) = \\
 &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log \left( \frac{p(c_i|z_i)}{p(c_i)} p(z_i) \right) + \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p(\mathbf{x}_i|z_i)) = \\
 &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p_{ij} \pi_j) + \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(f_j(\mathbf{x}_i; \boldsymbol{\theta}_j))
 \end{aligned} \tag{23}$$

and taking again the expectation of  $\log L_c$  yields the expression of the expected complete-data log-likelihood for WCA:

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^k) = \sum_i^N \sum_j^M \mathbb{E}_{\hat{\boldsymbol{\theta}}^k} \{z_{ij}\} \log(p_{ij} \pi_j) + \sum_i^N \sum_j^M \mathbb{E}_{\hat{\boldsymbol{\theta}}^k} \{z_{ij}\} \log(f_j(\mathbf{x}_i, \boldsymbol{\theta}_j)) \tag{24}$$

In the same manner, from Equation 19 with the definition 16:

$$\begin{aligned}
 CA, \text{ latent } c_i \rightarrow L_c &= \prod_{i=1}^N \prod_{j=1}^M \prod_{l=1}^L \left[ \sum_{c_i} p(c_i) p(z_i|c_i) \right] p(\mathbf{x}_i|z_i)^{z_{ij}} \Rightarrow \\
 \log L_c &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log \left( \sum_{l=1}^L p(c_i) p(z_i|c_i) \right) p(\mathbf{x}_i|z_i) = \\
 &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log \left( \sum_{l=1}^L p(c_i) p(z_i|c_i) \right) + \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p(\mathbf{x}_i|z_i)) = \\
 &= \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log p_{ij} + \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(f_j(\mathbf{x}_i; \boldsymbol{\theta}_j))
 \end{aligned} \tag{25}$$

and the same form of CA complete-data log-likelihood is derived as for the case of observed  $c_i$  in Equation 22. Expressions 22 and 24 can be also thought to emerge from Equations 13–15 through  $\prod_{i=1}^N \prod_{j=1}^M [p(\mathbf{x}_i, z_i)]^{z_{ij}}$  (latent  $c_i$ ) or  $\prod_{i=1}^N \prod_{j=1}^M [p(\mathbf{x}_i, z_i|c_i)]^{z_{ij}}$  (observed  $c_i$ ).

To complete the derivations in Table 1, the formulations for the E-step are based on the extraction of distributions  $p(z_i|\mathbf{x}_i)$  (latent  $c_i$ ) or  $p(z_i|\mathbf{x}_i, c_i)$  (observed  $c_i$ ), following from simple manipulations from the joint distributions in Equations 6 and 7 and replacing the definitions 16–18:

$$\begin{aligned}
 CA, \text{ observed } c_i \rightarrow p(z_i|\mathbf{x}_i, c_i) &= \frac{p(z_i, \mathbf{x}_i|c_i)}{p(\mathbf{x}_i|c_i)} \stackrel{(13)}{=} \frac{p(z_i|c_i)p(\mathbf{x}_i|z_i)}{\sum_{z_i} p(z_i|c_i)p(\mathbf{x}_i|z_i)} \Rightarrow \\
 \mathbb{E}_{\hat{\boldsymbol{\theta}}}\{z_{ij}\} &= p(z_i = j|\mathbf{x}_i, c_i) = \frac{p_{ij} f_j(\mathbf{x}_i; \boldsymbol{\theta}_j)}{\sum_{m=1}^M p_{im} f_m(\mathbf{x}_i; \boldsymbol{\theta}_m)}
 \end{aligned} \tag{26}$$

Similarly, for the WCA case:

$$\begin{aligned}
 WCA, \text{ observed } c_i \rightarrow p(z_i|\mathbf{x}_i, c_i) &= \frac{p(z_i, \mathbf{x}_i|c_i)}{p(\mathbf{x}_i|c_i)} \stackrel{(14)}{=} \frac{\frac{p(c_i|z_i)}{p(c_i)} p(z_i) p(\mathbf{x}_i|z_i)}{\sum_{z_i} \frac{p(c_i|z_i)}{p(c_i)} p(z_i) p(\mathbf{x}_i|z_i)} \Rightarrow \\
 \mathbb{E}_{\hat{\boldsymbol{\theta}}}\{z_{ij}\} &= p(z_i = j|\mathbf{x}_i, c_i) = \frac{p_{ij} \pi_j f_j(\mathbf{x}_i; \boldsymbol{\theta}_j)}{\sum_{m=1}^M p_{im} \pi_m f_m(\mathbf{x}_i; \boldsymbol{\theta}_m)}
 \end{aligned} \tag{27}$$

and, finally for  $CA$  with latent  $c_i$ :

$$\begin{aligned}
 CA, \text{ latent } c_i \rightarrow p(z_i|\mathbf{x}_i) &= \frac{p(z_i, \mathbf{x}_i)}{p(\mathbf{x}_i)} \stackrel{(15)}{=} \frac{[\sum_{c_i} p(z_i|c_i)p(c_i)]p(\mathbf{x}_i|z_i)}{\sum_{z_i} [\sum_{c_i} p(z_i|c_i)p(c_i)]p(\mathbf{x}_i|z_i)} \Rightarrow \\
 \mathbb{E}_{\hat{\theta}}\{z_{ij}\} &= p(z_i = j|\mathbf{x}_i) = \frac{p_{ij}f_j(\mathbf{x}_i; \theta_j)}{\sum_{m=1}^M p_{im}f_m(\mathbf{x}_i; \theta_m)}
 \end{aligned} \tag{28}$$

where it holds that  $\mathbb{E}_{\hat{\theta}}\{z_{ij}\} = 1 * p(z_i|\mathbf{x}_i) + 0 * p(z_i \neq j|\mathbf{x}_i) = p(z_i|\mathbf{x}_i)$  and similarly when  $c_i$  is observed.

All quantities in Table 1 (as well as the expected complete-data log-likelihoods) can be thus justified by means of the graphical models of Figure 1. It should be underlined that the probabilistic labels of  $WCA$  in definition 18 do not fulfill the requirement of being a probability distribution over  $z_i$  ( $\sum_{j=1}^M p_{ij} = 1, \forall i$  and  $p_{ij} \leq 1$ ). However, the normalized labels  $p'_{ij} = p_{ij} / \sum_{j=1}^M p_{ij}$  fulfill the above requirement and require no extra information to be computed. It is easy to see that this transformation leaves the E-step unaffected and simply leverages the log-likelihood functions by a constant term, thus leaving the parameter estimation properties unaffected.

We proceed by proving that  $CA$  EM estimation will monotonically increase the respective incomplete-data log-likelihood  $\log L$  at each EM iteration. We follow the same arguments as in McLachlan and Krishnan (2008, chap. 3.2). A similar proof for the  $WCA$  case is given in Côme et al. (2009).

**Proof** The incomplete-data log-likelihood  $\log L$  of Equation 19 equals the difference of the complete-data log-likelihood and the log-likelihood of the missing data  $z_i$  given the observed data  $\mathbf{x}_i$  (and  $c_i$  if it is observed)  $p(z_i|\mathbf{x}_i)$  or  $p(z_i|\mathbf{x}_i, c_i)$ . In both cases, from Equations 26 and 28, the latter term takes the form  $\sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(\frac{p_{ij}f_j(\mathbf{x}_i|\theta_j)}{\sum_{m=1}^M p_{im}f_m(\mathbf{x}_i|\theta_m)})$ , so from this result and Equation 21:

$$\begin{aligned}
 \log L &= \log L_c - \log(p(Z|X)) = \\
 \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(p_{ij}f_j(\mathbf{x}_i|\theta_j)) &- \sum_{i=1}^N \sum_{j=1}^M z_{ij} \log(\frac{p_{ij}f_j(\mathbf{x}_i|\hat{\theta}_j)}{\sum_{m=1}^M p_{im}f_m(\mathbf{x}_i|\hat{\theta}_m)})
 \end{aligned} \tag{29}$$

The relation holds for the expectations of these two terms, thus:

$$\begin{aligned}
 \log L &= \mathbb{E}_{\hat{\theta}}\{\log L_c\} - \mathbb{E}_{\hat{\theta}}\{\log(p(Z|X))\} = \\
 \underbrace{\sum_{i=1}^N \sum_{j=1}^M \mathbb{E}_{\hat{\theta}}\{z_{ij}\} \log(p_{ij}f_j(\mathbf{x}_i|\theta_j))}_{Q(\theta, \hat{\theta})} &- \underbrace{\sum_{i=1}^N \sum_{j=1}^M \mathbb{E}_{\hat{\theta}}\{z_{ij}\} \log(\frac{p_{ij}f_j(\mathbf{x}_i|\hat{\theta}_j)}{\sum_{m=1}^M p_{im}f_m(\mathbf{x}_i|\hat{\theta}_m)})}_{H(\theta, \hat{\theta})}
 \end{aligned} \tag{30}$$

Note, that since the expectation is as in Equation 26, the term  $H(\theta, \hat{\theta})$  is the entropy of the latent data. Between two iterations of the algorithm, the difference of the incomplete-data log-likelihood (where the result of the respective M-step replaces  $\theta$ )  $\log L(\hat{\theta}^{k+1}) - \log L(\hat{\theta}^k)$  will be:

$$\begin{aligned}
 \log L(\hat{\theta}^{k+1}) - \log L(\hat{\theta}^k) &= \\
 \underbrace{\{Q(\hat{\theta}^{k+1}, \hat{\theta}^k) - Q(\hat{\theta}^k, \hat{\theta}^k)\}}_A &- \underbrace{\{H(\hat{\theta}^{k+1}, \hat{\theta}^k) - H(\hat{\theta}^k, \hat{\theta}^k)\}}_B
 \end{aligned} \tag{31}$$

As a result of analytically maximizing  $Q(\theta, \hat{\theta})$  at the M-step, it holds that  $A \geq 0$ , where equality will hold when the previous estimate  $\hat{\theta}$  is already a maximum of  $Q$ . It remains to show that  $B \leq 0$ , so that  $-B \geq 0$  and the difference of the consecutive incomplete-data log-likelihood values will be

positive (or 0):

$$\begin{aligned}
 B &= H(\hat{\boldsymbol{\theta}}^{k+1}, \hat{\boldsymbol{\theta}}^k) - H(\hat{\boldsymbol{\theta}}^k, \hat{\boldsymbol{\theta}}^k) \stackrel{(26,28)}{=} \\
 &\sum_{i=1}^N \sum_{j=1}^M \underbrace{\frac{p_{ij} f_j(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_j^k)}{\sum_{m=1}^M p_{im} f_m(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_m^k)}}_{t_{ij}^k} \log\left(\frac{p_{ij} f_j(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_j^{k+1})}{\sum_{m=1}^M p_{im} f_m(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_m^{k+1})}\right) - \\
 &\sum_{i=1}^N \sum_{j=1}^M \frac{p_{ij} f_j(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_j^k)}{\sum_{m=1}^M p_{im} f_m(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_m^k)} \log\left(\frac{p_{ij} f_j(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_j^k)}{\sum_{m=1}^M p_{im} f_m(\mathbf{x}_i | \hat{\boldsymbol{\theta}}_m^k)}\right) = \\
 &\sum_{i=1}^N \sum_{j=1}^M t_{ij}^k \log(t_{ij}^{k+1}) - \sum_{i=1}^N \sum_{j=1}^M t_{ij}^k \log(t_{ij}^k) = \sum_{i=1}^N \sum_{j=1}^M t_{ij}^k \log\left(\frac{t_{ij}^{k+1}}{t_{ij}^k}\right) \leq 0
 \end{aligned} \tag{32}$$

where the last result is due to Gibb's inequality. Hence  $-B \geq 0$  and the proof is concluded.  $\blacksquare$

## Appendix B. Relation to Posterior Regularization

Ganchev et al. (2010, Section 4) have shown that all methods so far proposed to exploit side-information in the form of generic constraints (see Section 2) are equivalent under certain assumptions. It hence suffices to discuss the relation of our framework to that of Posterior Regularization (PR). For the purpose of comparing these algorithms to PR, it is beneficial to formulate the same class of algorithms from the alternative, maximization–maximization viewpoint (Bishop, 2006, Chap. 9.4). In this case, the marginal log-likelihood  $\log p(X|\boldsymbol{\theta})$ , is decomposed as:  $\log p(X|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$ , where,  $\mathcal{L}(q, \boldsymbol{\theta}) = \sum_Z q(Z) \log\left\{\frac{p(X, Z|\boldsymbol{\theta})}{q(Z)}\right\}$  a functional of some distribution  $q(Z)$  and  $KL(q||p) = -\sum_Z q(Z) \log\left\{\frac{p(Z|X, \boldsymbol{\theta})}{q(Z)}\right\}$  a Kullback–Leibler divergence between  $q(Z)$  and the posterior distribution  $p = p(Z|X, \boldsymbol{\theta})$ . The EM algorithm then consists of two consecutive maximization steps. First, the term  $KL(q||p)$  is maximized over  $q$ , resulting in  $q(Z) = p(Z|X, \boldsymbol{\theta})$  (KL divergence is zero when  $q = p$ ). Then, the functional  $\mathcal{L}(q, \boldsymbol{\theta})$  is maximized over  $\boldsymbol{\theta}$  after having replaced the  $q(Z)$  found in the first step, providing a new estimate  $\hat{\boldsymbol{\theta}}$ . This alternative EM presentation will lead to the exact same formulations of our algorithms summarized in Table 1.

On the contrary, PR proceeds by constraining the distribution  $q(Z)$  so that expectations  $\mathbb{E}_q\{\boldsymbol{\phi}(X, Z)\}$  of constraint features external to the model,  $\boldsymbol{\phi}(X, Z)$ , are respected. Formally, the constraint posterior set  $Q = \{q(Z) : \mathbb{E}_q\{\boldsymbol{\phi}(X, Z)\} \leq \mathbf{b}\}$  is imposed. In order to derive one of the proposed algorithms in the PR setting, one has to derive the same  $q(Z)$  emerging in our own corresponding formulation (E-step). Noting the distribution of the derived probabilistic labels as  $P'$ , the *DCA* algorithm is derived in the PR framework by directly restricting  $Q = P'$ . For *CA* and *WCA*, one might note that the final  $q(Z)$  could be derived by maximizing in the first step the sum  $KL(q||p(Z|X)) + KL(q||p')$  over  $q$ , i.e., by augmenting the objective function with an extra Kullback–Leibler divergence term related to the contextual posteriors. More specifically, this augmentation can be achieved by considering instance-wise, external constraint features  $\boldsymbol{\phi}(X, Z)$  that enforce agreement between the model-based posteriors  $p(Z|X)$  (what has been hereby called *bottom-up* information) and context-based (*top-down*) posteriors  $\mathbf{p}_i$ . To this end, the same definition of constraint features used in Equation 20 and the objective definition of Equation 3 in Ganchev et al. (2010) can be used<sup>13</sup>. For the case of *CA*, it should be additionally assumed that the model-based priors are uniform, which effectively cancels out the mixing coefficients  $\pi_j$  in the resulting  $q(Z)$  distribution for this algorithm.

13. Concerning this objective, we set  $\sigma = 1$  and replace the norm  $\|\cdot\|_\beta$  with KL divergence. As thereby noted by the authors, this replacement is allowed since KL divergence is a convex penalty function.

All algorithms proposed here can be hence also formulated in the PR framework. Such practical equivalence notwithstanding, the two frameworks are not conceptually interchangeable, with several reasons contending in favour of the one proposed here. First, the two basic features distinguishing our framework from PR, namely, the strictly probabilistic representation of contextual information and the direct embedding of side-information into a graphical model’s structure, offer a very intuitive and straightforward modeling perspective. This is clear in the exemplary application provided. Conversely, we argue that passing through unintuitive definitions of complex external features can render the applicability of PR less straightforward for many applications. Second, our framework brings naturally forward two different algorithms (each depending in a different structure of the graph) using the standard EM methodology. In order to get the same formulations in PR, further unintuitive assumptions for  $CA$  are needed. Third, the optimization procedure in our case follows standard, simple reasoning and avoids sophisticated optimization tools like Langrangian duality. Fourth, as noted by Ganchev et al. (2010), in the PR derivation it is evident that the standard model likelihood is traded-off with satisfaction of constraint feature expectations; yet, in our own derivation, it is explicitly shown that an augmented model’s likelihood is exactly optimized<sup>14</sup>. Last, but not least, our formulation makes the applicability of the missing information principle more profound.

Concluding, PR and similar frameworks are more generic devices, allowing one to exploit types of side-information beyond the specific limitations imposed here. However, our own framework is comparatively more intuitive and parsimonious from both modeling and derivation perspectives, and, thus, easier applicable in a wide range of applications, where adopting PR can be viewed as an unnecessary complexification.

### Appendix C. Effects of contextual negentropy on observed information matrices of finite mixture models

We hereby formally substantiate the effects observed in Section 4.2 for the general case of any FMM. In the following proofs we assume that only contextual negentropy  $NE$  can be varied and that all context-aware methods converge to the same MLE  $\hat{\theta}$ , applied on the same dataset  $X$ . In Section 4.1 we have shown that given “correct” context, identical estimation problems with different  $NE$  will indeed have a common local maximum very close to the supervised MLE.

**Proposition** *Let  $I_c$  be the expected complete-data observed information matrix of any finite mixture model with parameters  $\theta$  evaluated at  $\hat{\theta}$  for any of the context-aware EM algorithms in Table 1 with contextual negentropy  $NE$ . For given observed data  $X$  of cardinality  $N$ , matrices  $I_c$  for different  $NE$  will be equal.*

**Proof** In the general case of a FMM of  $M$  arbitrary mixtures, the estimated parameters consist of  $M - 1$  mixture coefficients  $\pi_j, j \in [1, M - 1]$  and  $M \times P$  internal mixture parameters  $\theta_p^k, p \in [1, P], k \in [1, M]$ , where superscript  $k$  reflects that a parameter belongs to the  $k^{th}$  mixture, yielding a total of  $W = M \times (P + 1) - 1$  estimated parameters  $\theta_l^k, l \in [1, W]$ .

Matrix  $I_c$  is a  $W \times W$  symmetric matrix, where each element  $I_c^{l_1, l_2}, l_1, l_2 \in [1, W]$  is derived by definition as  $I_c^{l_1, l_2} = \mathbb{E}_{\hat{\theta}} \left\{ -\frac{\partial^2 \log L_c}{\partial \theta_{l_1} \partial \theta_{l_2}} \right\}$  where  $\log L_c$  the complete-data log-likelihood of the respective estimation method. We study separately the cases that can occur for any given combination  $\theta_{l_1}, \theta_{l_2}$ . Recall that different  $NE$  levels only affect the E-step and thus operate on the information matrix only by altering the expectations  $\mathbb{E}_{\hat{\theta}} \{z_{ij}\}$ , which from now on are denoted as  $\bar{z}_{ij}$ .

For any FMM with  $M$  mixtures, it holds through the definition that:

$$I_c^{\pi_j, \pi_q} = \begin{cases} \frac{\sum_{i=1}^N \bar{z}_{ij}}{\hat{\pi}_j^2} + \frac{N - \sum_{i=1}^N \sum_{k=1}^{M-1} \bar{z}_{ik}}{(1 - \sum_{k=1}^{M-1} \hat{\pi}_k)^2} & , j = q \\ \frac{N - \sum_{i=1}^N \sum_{k=1}^{M-1} \bar{z}_{ik}}{(1 - \sum_{k=1}^{M-1} \hat{\pi}_k)^2} & , j \neq q \end{cases} \quad (33)$$

14. Section 4.1 shows that the “distortion” of the  $CA, WCA$  likelihoods (compared to  $US$ ) is favourable.

Since at the common evaluation point  $\hat{\boldsymbol{\theta}}$  it holds from the M-step that  $\hat{\pi}_j = \frac{\sum_{i=1}^N \bar{z}_{ij}}{N}, \forall j$ , for any  $NE$ , replacing the quantity  $\sum_{i=1}^N \bar{z}_{ij} = N\hat{\pi}_j, \forall j$  in Equation 33, the evaluation of  $I_c^{\pi_j, \pi_q}$  yields:

$$I_c^{\pi_j, \pi_q} = \begin{cases} \frac{N}{\hat{\pi}_j} + \frac{N}{1 - \sum_{k=1}^{M-1} \hat{\pi}_k} & , j = q \\ \frac{N}{1 - \sum_{k=1}^{M-1} \hat{\pi}_k} & , j \neq q \end{cases} \quad (34)$$

Furthermore, all elements of matrix  $I_c$  of the form  $I_c^{\pi_j, \theta_p^k}, \forall j, p, k$  or  $I_c^{\theta_p^m, \theta_p^k}, \forall p, k \neq m$  will be 0, since they appear in separate linear terms of  $\log L_c$ . It is thus shown that the majority of elements of  $I_c$  will be constant or always 0 irrespectively of  $NE$  (thus also constant). Regarding the  $I_c$  elements related to parameters belonging in the same mixture  $k$ ,  $I_c^{\theta_p^k, \theta_q^k}$ , either for diagonal elements ( $p = q$ ) or non-diagonal ( $p \neq q$ ), the general formulation as derived by the definition is:

$$I_c^{\theta_p^k, \theta_q^k} = \sum_{i=1}^N \bar{z}_{ik} \frac{\partial^2 \log f_k(\mathbf{x}_i; \boldsymbol{\theta}^k)}{\partial \theta_p^k \partial \theta_q^k} \quad (35)$$

Depending on the particular type of the mixture's probability density function form and the parameters  $\theta_p^k, \theta_q^k$ , the quantities  $\frac{\partial^2 \log f_k(\mathbf{x}_i; \boldsymbol{\theta}^k)}{\partial \theta_p^k \partial \theta_q^k}$  can be either 0 (thus constant) or dependent only on some (or all) of the mixture's parameters and independent of the observed data  $X$ . In the latter case, the constant (upon evaluation at a common  $\hat{\boldsymbol{\theta}}$  for all  $NE$ ) across all  $i$ -s quantity can be moved outside the summation over  $i$  and the same argument as in Equation 33 can be employed to yield the same matrix element value irrespectively of  $NE$ . In the general case, however, this quantity will be non-zero and dependent on both the mixture's parameters and the observed data  $X$ . Note that every  $I_c$  matrix element (including the ones considered above) is a special condition of this general form.

For the general case, we can prove a more general result in terms of the expected rather than the observed information. The quantity in the sum of Equation 35 for some  $\mathbf{x}$  can be written as  $\frac{p(\mathbf{x}|z=k)}{p(\mathbf{x})} g(\mathbf{x}, \boldsymbol{\theta}^k)$ , where  $\bar{z}_k = p(z = k|\mathbf{x}) = \frac{p(\mathbf{x}|z=k)}{p(\mathbf{x})}$  and  $g(\mathbf{x}, \boldsymbol{\theta}^k) = \frac{\partial^2 \log f_k(\mathbf{x}_i; \boldsymbol{\theta}^k)}{\partial \theta_p^k \partial \theta_q^k}$ , hence the expected value of this function (of the observed random variable  $\mathbf{x}$ ) with respect to  $\mathbf{x}$  is:

$$\int_X \frac{p(\mathbf{x}|z=k)}{p(\mathbf{x})} g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \int_X g(\mathbf{x}) f_k(\mathbf{x}, \boldsymbol{\theta}^k) d\mathbf{x} = \mathbb{E}_{X|Z=k} \{g(\mathbf{x})\} = \text{const}. \quad (36)$$

where this expectation is independent of the posteriors  $p(z = k|\mathbf{x})$  (which is the sole entity affected by context) and constant as long as  $p(\mathbf{x}|z = k) = f_k(\mathbf{x}, \boldsymbol{\theta}^k)$  is the same for different  $NE$  values. The latter condition holds when all methods converge to the same MLE in general, as assumed here. Consequently, the observed information can be approximated through the expected information as  $N\mathbb{E}_{X|Z=k} \{g(\mathbf{x})\}$  given large enough  $N$ .  $\blacksquare$

We hereafter show that the missing information matrix  $I_m$  vanishes to the zero matrix as contextual negentropy  $NE$  approaches 1, and the probabilistic labels tend towards the ground truth data labels.

**Proposition** *Let  $I_m$  be the observed missing information matrix of any finite mixture model with parameters  $\boldsymbol{\theta}$  evaluated at  $\hat{\boldsymbol{\theta}}$  for any of the context-aware EM algorithms in Table 1 with contextual negentropy  $NE$ . For given observed data  $X$  of cardinality  $N$ , matrices  $I_m$  tend to the zero matrix as  $NE \rightarrow 1$ .*

**Proof** Matrix  $I_m$  is a  $W \times W$  symmetric matrix, where each element  $I_m^{l_1, l_2}, l_1, l_2 \in [1, W]$  is derived as  $I_m^{l_1, l_2} = \text{Cov}_{\hat{\boldsymbol{\theta}}} \left\{ \frac{\partial \log L_c}{\partial \theta_{l_1}}, \frac{\partial \log L_c}{\partial \theta_{l_2}} \right\}$  using one of alternative definitions, where  $\log L_c$  the complete-data log-likelihood of the respective estimation method as summarized in Table 1. Matrix  $I_m$  is thus

defined as the variance–covariance of the complete–data score (vector of first order derivatives of the complete–data log–likelihood). Note that matrices  $I_c$ ,  $I_m$  have the same structure for a given estimation problem, since the latter depends only on the number and type of parameters estimated.

We again study separately the cases that can occur for the combinations of  $\theta_{l_1}, \theta_{l_2}$ :

$$I_m^{\pi_j, \pi_j} = \text{Var}\left\{\frac{\sum_{i=1}^N z_{ij}}{\hat{\pi}_j} - \frac{\sum_{i=1}^N z_{iM}}{\hat{\pi}_M}\right\} = \frac{\sum_{i=1}^N \bar{z}_{ij}(1 - \bar{z}_{ij})}{\hat{\pi}_j^2} + \frac{2 \sum_{i=1}^N \bar{z}_{ij} \bar{z}_{iM}}{\hat{\pi}_j \hat{\pi}_M} + \frac{\sum_{i=1}^N \bar{z}_{iM}(1 - \bar{z}_{iM})}{\hat{\pi}_M^2} \quad (37)$$

$$I_m^{\pi_j, \pi_k} = \text{Cov}\left\{\frac{\sum_{i=1}^N z_{ij}}{\hat{\pi}_j} - \frac{\sum_{i=1}^N z_{iM}}{\hat{\pi}_M}, \frac{\sum_{i=1}^N z_{ik}}{\hat{\pi}_k} - \frac{\sum_{i=1}^N z_{iM}}{\hat{\pi}_M}\right\} = \frac{\sum_{i=1}^N \bar{z}_{ij} \bar{z}_{ik}}{\hat{\pi}_j \hat{\pi}_k} - \frac{\sum_{i=1}^N \bar{z}_{ij} \bar{z}_{iM}}{\hat{\pi}_j \hat{\pi}_M} - \frac{\sum_{i=1}^N \bar{z}_{ik} \bar{z}_{iM}}{\hat{\pi}_k \hat{\pi}_M} + \frac{\sum_{i=1}^N \bar{z}_{iM}(1 - \bar{z}_{iM})}{\hat{\pi}_M^2} \quad (38)$$

$$I_m^{\pi_j, \theta_p^k} = \text{Cov}\left\{\frac{\sum_{i=1}^N z_{ij}}{\hat{\pi}_j} - \frac{\sum_{i=1}^N z_{iM}}{\hat{\pi}_M}, \sum_{i=1}^N z_{ik} \frac{\partial}{\partial \theta_p^k}(\log f_k(\mathbf{x}_i; \theta^k))\right\} = \begin{cases} -\frac{\sum_{i=1}^N b_{k,p}^2(\mathbf{x}_i) \bar{z}_{ij} \bar{z}_{ik}}{\hat{\pi}_j} + \frac{\sum_{i=1}^N b_{k,p}^2(\mathbf{x}_i) \bar{z}_{iM} \bar{z}_{ik}}{\hat{\pi}_M} & , j \neq k \\ \frac{\sum_{i=1}^N b_{j,p}^2(\mathbf{x}_i) \bar{z}_{ij}(1 - \bar{z}_{ik})}{\hat{\pi}_j} + \frac{\sum_{i=1}^N b_{i,p}^2(\mathbf{x}_i) \bar{z}_{ij} \bar{z}_{iM}}{\hat{\pi}_M} & , j = k \end{cases} \quad (39)$$

$$I_m^{\theta_p^j, \theta_q^k} = \text{Cov}\left\{\sum_{i=1}^N z_{ij} \frac{\partial}{\partial \theta_p^j}(\log f_j(\mathbf{x}_i; \theta^j)), \sum_{i=1}^N z_{ik} \frac{\partial}{\partial \theta_q^k}(\log f_k(\mathbf{x}_i; \theta^k))\right\} = \begin{cases} -\sum_{i=1}^N b_{j,p}^2(\mathbf{x}_i) b_{k,q}^2(\mathbf{x}_i) \bar{z}_{ij} \bar{z}_{ik} & , j \neq k, \forall p, q \\ \sum_{i=1}^N b_{j,p}^2(\mathbf{x}_i) b_{j,q}^2(\mathbf{x}_i) \bar{z}_{ij}(1 - \bar{z}_{ij}) & , j = k, \forall p, q \end{cases} \quad (40)$$

where we have replaced for brevity:  $b_{j,p}(\mathbf{x}_i) = \frac{\partial}{\partial \theta_p^j}(\log f_j(\mathbf{x}_i; \theta^j))$ . Expressions 37–40 follow directly from the definition through simple manipulations, employing known properties of the variance and covariance operators—specifically, the properties concerning the (co)variance of sums and linear combinations of random variables and the definitions of (co)variance in terms of expectations—as well as the following facts that hold for random indicator variables  $z_{ij}$ : (a)  $\text{Var}\{z_{ij}, z_{lk}\} = \text{Cov}\{z_{ij}, z_{lk}\} = 0, \forall k, j \text{ if } l \neq i$  because of the iid assumption in FMMs. (b)  $\mathbb{E}\{z_{ij}^2\} = \mathbb{E}\{z_{ij}\}$ , because  $z_{ij}^2 = z_{ij}$ , since  $z_{ij} \in \{0, 1\}$  and (c)  $\mathbb{E}\{z_{ij} z_{ik}\} = 0 \text{ if } j \neq k$ , because  $z_{ij} z_{ik} = 0$  for  $j \neq k$  since when  $z_{ij} = 1 \Rightarrow z_{ik} = 0$  and vice-versa.

Note that all matrix elements of  $I_m$  are made of terms containing summations of the form  $\bar{z}_{ij} \bar{z}_{ik}$  or  $\bar{z}_{ij}(1 - \bar{z}_{ij})$ . At  $NE = 1$  (where the problem reduces to supervised learning since probabilistic labels become deterministic) it holds that  $\bar{z}_{ij} = 1$  for some  $j$  and  $\bar{z}_{ik} = 0, \forall k \neq j$ , thus also  $1 - \bar{z}_{ij} = 0$ . Consequently, all summations and, as a result, all elements of matrix  $I_m$  will be 0 at  $NE = 1$ .

Furthermore, it is easy to see that, with evaluation at the same  $\hat{\theta}$  for the same dataset  $X$  which renders quantities  $b_{j,p}(\mathbf{x}_i)$  and  $f_j(\mathbf{x}_i; \theta^j)$  invariant to  $NE$ , products  $\bar{z}_{ij} \bar{z}_{ik}$  and  $\bar{z}_{ij}(1 - \bar{z}_{ij})$  degrade as  $NE$  is increased (recall that increasing  $NE$ —with “correct” context—yields higher value of probabilistic label  $p_{ij}$  if  $y_j = 1$  and lower otherwise). That holds through the definition of the E–step for context–aware methods, because  $\bar{z}_{ij}^1 < \bar{z}_{ij}^2, 1 - \bar{z}_{ij}^1 > 1 - \bar{z}_{ij}^2$  if  $y_j = 1$  and  $\bar{z}_{ik}^1 \geq \bar{z}_{ik}^2$  if  $y_k = 0$ , for  $NE_1 < NE_2$ <sup>15</sup>. As a result, the absolute value of all  $I_m$  matrix elements degrades with increasing

15. These statements hold exactly for the *WCA* case. For the *CA* method, it only holds for samples  $\mathbf{x}_i$  where  $p_{ij} > \hat{\pi}_j$  when  $y_j = 1$ . However, it is reasonable to expect that, especially as  $NE$  is increased, the probabilistic label is more informative than the prior  $\hat{\pi}_j$ . Furthermore, this condition holds exactly when uninformative (uniform) priors are employed.

$NE^{16}$ . ■

## Appendix D. Generation of actual and initial FMM distributions for simulations with artificial data

**Scenarios involving mixtures of univariate normal distributions.** *Actual distributions:* Mean  $\mu_1 \in [0, 1]$ , standard deviation for mixture  $j$ ,  $s_j \in [0.1, 0.6]$ , means  $\mu_{2,3}$  analytically computed so that the corresponding mixture  $j = 2, 3$  exhibits separability with mixture  $j - 1$  of  $SKL \in [0.1, 3]$  (two-mixture problems) or  $SKL \in [3, 20]$  (three-mixture problems). *Initialization:* standard deviation initialization for each mixture  $j$ ,  $\hat{s}_j^0 \in [0.1, 0.6]$ , mean initialization for mixture  $j$ ,  $\hat{\mu}_j^0$ , analytically computed so that the corresponding initial mixture  $j$  exhibits separability with the actual mixture  $j$  of  $IKL \in [0.1, 3]$  (two-mixture problems) or  $IKL \in [0.1, 1]$  (three-mixture problems).

**Scenario involving two mixtures of multivariate normal distributions.** *Actual distributions:* elements  $k$  of mean vector  $\boldsymbol{\mu}_1$ ,  $\mu_{1k} \in [0, 1]$ , elements  $k$  of mean vector  $\boldsymbol{\mu}_2$ ,  $\mu_{2k}$ , analytically computed so that mixture  $j = 2$  exhibits separability with mixture  $j = 1$  of  $SKL \in [0.1, 3]$ , covariance matrix for mixture  $j$ ,  $\Sigma_j$ , as random symmetric matrix with all eigenvalues  $\lambda_k$  drawn from uniform distributions  $\lambda_k \in [0, 0.5]$  (positive semi-definite). *Initialization:* covariance matrix for mixture  $j$ ,  $\hat{\Sigma}_j^0$  as random symmetric matrix with all eigenvalues  $\lambda_k$  drawn from uniform distributions  $\lambda_k \in [0, 0.5]$ , elements  $k$  of mean vector for mixture  $j$ ,  $\hat{\boldsymbol{\mu}}_j^0$ ,  $\hat{\mu}_{jk}$  analytically computed so that the corresponding initial mixture exhibits separability with actual mixture  $j$ ,  $IKL \in [0.1, 3]$ .

**Scenario involving two mixtures of Maxwell–Boltzmann distributions.** *Actual distributions:* Distribution parameter  $\alpha_1 \in [1, 6]$ , distribution parameter  $\alpha_2$  analytically computed so that the corresponding mixture exhibits separability with mixture  $j = 1$ ,  $SKL \in [0.1, 3]$ . *Initialization:* Distribution parameter  $\hat{\alpha}_j$  of mixture  $j$ , analytically computed so that the corresponding initial mixture exhibits separability with the actual mixture  $j$ ,  $IKL \in [0.1, 3]$ .

**Scenario involving two mixtures of univariate, first-order linear regression models.** *Actual distributions:* Zero-order regression coefficient (intercept) of mixture  $j$ ,  $\beta_{j0} \in [-1, 1]$ , first-order regression coefficient (slope) of mixture  $j$ ,  $\beta_{j1} = \tan(\theta_j)$ , where  $\theta_j \in [-\frac{\pi}{3}, \frac{\pi}{3}]$ , error term (noise) of mixture  $j$ ,  $\epsilon_j \in [0.5, 2]$ . *Initialization:* Identically to actual distributions, no “separability” safeguards taken.

## References

- Kartik Audhkhasi and Shrikanth S. Narayanan. A Globally-Variant Locally-Constant Model for Fusion of Labels from Multiple Diverse Experts without Using Reference Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):769–783, 2013.
- Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Semi-supervised Clustering by Seeding. In *Proceedings of the Nineteenth International Conference on Machine Learning, ICML ’02*, pages 27–34, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.
- Kedar Bellare, Gregory Druck, and Andrew McCallum. Alternating projections for learning with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI ’09*, pages 43–50, Arlington, Virginia, United States, 2009. AUAI Press.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

---

16. This statement is straightforward for all expressions in Equations 37– 40 where the consisting terms have the same sign. It can be shown that it also holds for expressions in Equation 38 and the first leg of Equation 39 due to the different rates at which the signed terms vanish with increasing  $NE$ .

- Charles Bouveyron and Stéphane Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, November 2009.
- Nicholas Bryan and Gautham Mysore. An efficient posterior regularized latent variable model for interactive sound source separation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 208–216. JMLR Workshop and Conference Proceedings, May 2013.
- Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10*, pages 101–110, New York, NY, USA, 2010. ACM.
- Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding Semi-Supervision with Constraint-Driven Learning. In *ACL*, pages 280–287, Prague, Czech Republic, 6 2007. Association for Computational Linguistics.
- Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
- Harr Chen, Edward Benson, Tahira Naseem, and Regina Barzilay. In-domain relation discovery with meta-constraints via posterior regularization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, pages 530–540, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- Yi-Chen Chen, Vishal M. Patel, Jaishanker K. Pillai, Rama Chellappa, and P. Jonathon Phillips. Dictionary learning from ambiguously labeled data. In *CVPR*, pages 353–360. IEEE, 2013.
- Etienne Côme, Latifa Oukhellou, Thierry Denœux, and Patrice Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334–348, 2009.
- Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, July 2011.
- Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. Pac generalization bounds for co-training. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 375–382. MIT Press, 2002.
- Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.
- Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, pages 595–602, New York, NY, USA, 2008. ACM.
- Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2nd edition, 2001.
- James Foulds and Padhraic Smyth. Multi-instance mixture models and semi-supervised learning. In *Proceedings of the SIAM International Conference on Data Mining*, 2011.
- Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, August 2010.

- Benjamin Georgi. *Context Specific Independence Mixture Models for Cluster Analysis of Biological Data*. 2009.
- Soumya Ghosh, Soundararajan Srinivasan, and Burton Andrews. Using weak supervision in learning gaussian mixture models. In *Proceedings of International Joint Conference on Neural Networks, 2009*, pages 973–979, 2009.
- Luheng He, Jennifer Gillenwater, and Ben Taskar. Graph-Based Posterior Regularization for Semi-Supervised Structured Prediction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 38–46. Association for Computational Linguistics, 2013.
- Armand Joulin and Francis R. Bach. A convex relaxation for weakly supervised classifiers. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1279–1286. ACM, 2012.
- Pieter-Jan Kindermans, Hannes Verschore, David Verstraeten, and Benjamin Schrauwen. A P300 BCI for the masses: prior information enables instant unsupervised spelling. In *Proceedings of Advances in Neural Information Processing Systems 25*, page 9, 2012a.
- Pieter-Jan Kindermans, David Verstraeten, and Benjamin Schrauwen. A Bayesian Model for Exploiting Application Constraints to Enable Unsupervised Training of a P300-based BCI. *PLoS ONE*, 7(4):337–358, 2012b.
- Balaji Lakshminarayanan and Yee Whye Teh. Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *ArXiv e-prints*, April 2013.
- Robert Leeb, Serafeim Perdikis, Luca Tonin, Andrea Biasiucci, Michele Tavella, Marco Creatura, Alberto Molina, Abdul Al-Khodairy, Tom Carlson, and José d. R. Millán. Transferring brain-computer interfaces beyond the laboratory: Successful application control for motor-disabled users. *Artificial Intelligence in Medicine*, 59(2):121–132, 2013.
- E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 1998.
- Percy Liang, Michael I. Jordan, and Dan Klein. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pages 641–648, New York, NY, USA, 2009. ACM.
- Chao Liu and Yi-Min Wang. TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 225–232, New York, NY, USA, 2012. ACM.
- Jie Luo and Francesco Orabona. Learning from candidate labeling sets. In *Proceedings of Advances in Neural Information Processing Systems 23*, pages 1504–1512. 2010.
- Gideon S. Mann and Andrew McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, pages 593–600, New York, NY, USA, 2007. ACM.
- Gideon S. Mann and Andrew McCallum. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984, March 2010.
- Andrew McCallum and Kamal Nigam. Text Classification by Bootstrapping with Keywords, EM and Shrinkage. In *Workshop On Unsupervised Learning In Natural Language Processing*, pages 52–58, 1999.

- Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2nd edition, 2008.
- Xiao-Li Meng and Donald B. Rubin. On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra and its Applications*, 199, Supplement 1(0):413–425, 1994.
- José d. R. Millán. On the need for on-line learning in brain-computer interfaces. In *Proceedings of the IEEE International Joint Conference on Neural Networks, 2004*, volume 4, pages 2877–2882, 2004.
- Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification with auxiliary probabilistic information. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM), 2011*, pages 477–486, 2011.
- Terence Orchard and Max A. Woodbury. A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 697–715. University of California Press, 1972.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.
- Serafeim Perdikis, Robert Leeb, John Williamson, Andrew Ramsay, Michele Tavella, Lorenzo Desideri, Evert-Jan Hoogerwerf, Abdul Al-Khodairy, Roderick Murray-Smith, and José d. R. Millán. Clinical evaluation of BrainTree, a motor imagery hybrid BCI speller. *Journal of Neural Engineering*, 11(3):036003, 2014.
- Novi Quadrianto, James Petterson, and Alex J. Smola. Distribution matching for transduction. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1500–1508. Curran Associates, Inc., 2009.
- Sundararajan Sellamanickam, Charu Tiwari, and Sathiya K. Selvaraj. Regularized Structured Output Learning with Partial Labels. In *SDM*, pages 1059–1070. SIAM / Omnipress, 2012.
- Noam Shental, Aharon Bar-Hillel, Tomer Hertz, and Daphna Weinshall. Computing Gaussian mixture models with EM using equivalence constraints. In *Proceedings of Advances in Neural Information Processing Systems 16*, 2004.
- Yuyin Sun, Yin Zhang, and Zhihua Zhou. Multi-label learning with weak label. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010.
- Richard .S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford book. MIT Press, 1998. ISBN 9780262193986. URL <http://books.google.ch/books?id=CAFR6IBF4xYC>.
- Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.
- Ruth Urner, Shai Ben-David, and Ohad Shamir. Learning from weak teachers. *Journal of Machine Learning Research - Proceedings Track*, 22:1252–1260, 2012.
- Bishan Yang and Claire Cardie. Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 325–335. Association for Computational Linguistics, 2014.

- Yutaka Yasui, Margaret Pepe, Li Hsu, Bao-Ling Adam, and Ziding Feng. Partially Supervised Learning Using an EM-Boosting Algorithm. *Biometrics*, 60(1):199–206, 2004.
- Ping Zhang and Zoran Obradovic. Learning from inconsistent and unreliable annotators by a gaussian mixture model and bayesian information criterion. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases - Volume Part III*, ECML PKDD'11, pages 553–568. Springer-Verlag, 2011.
- Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15(1):1799–1847, January 2014.