NEURAL VARIATIONAL IDENTIFICATION AND FILTERING FOR STOCHASTIC NON-LINEAR DYNAMICAL SYSTEMS WITH APPLICATION TO NON-INTRUSIVE LOAD MONITORING

Henning Lange, Mario Bergés, Zico Kolter

Carnegie Mellon University

ABSTRACT

In this paper, an algorithm for performing System Identification and inference of the filtering recursion for stochastic non-linear dynamical systems is introduced. Additionally, the algorithm allows for enforcing domain-constraints of the state variable. The algorithm makes use of an approximate inference technique called Variational Inference in conjunction with Deep Neural Networks as the optimization engine. Although general in its nature, the algorithm is evaluated in the context of Non-Intrusive Load Monitoring, the problem of inferring the operational state of individual electrical appliances given aggregate measurements of electrical power collected in a home.

Index Terms— Dynamical Systems, Variational Inference, Deep Learning, System Identification, Filtering

1. INTRODUCTION

System identification and inference for dynamical systems has a long history [1]. In this paper we consider systems relating observations $x \in \mathbb{R}^{T\times M}$ to unknown latent states $z \in \mathbb{R}^{T\times C}$, with a joint distribution $p_\Theta(x, y)$, known up to parameter Θ, that can be factorized as follows:

$$z_t \sim p_\Theta(z_t|z_{t-1}) \quad x_t \sim p_\Theta(x_t|z_t) \tag{1}$$

When the dynamics are linear, optimal solutions such as the Kalman filters [2] and subspace methods [3] exist. However, when the system dynamics are non-linear and stochastic, non-optimal techniques such as particle filters in conjunction with Expectation Maximization need to be resorted to. The bottleneck for these approaches is oftentimes the computation of the filtering recursion $p_\Theta(z_t|x_{1:t})$, which for many latent variable models is computationally intractable. We propose a novel algorithm in which maximizing the data likelihood $p_\Theta(x_{1:T})$ is performed jointly alongside approximating inference of the intractable filtering recursion. The algorithm makes use of an approximate statistical inference technique called Variational Inference (VI), which has recently received increased attention from the Machine Learning community. Specifically, recent breakthroughs have improved VI's applicability [4], scalability [5, 6] and accuracy [7, 8]. See [9, 10] for reviews of the approach.

The algorithm will be showcased in the context of Non-Intrusive Load Monitoring [11] (NILM), which is the problem of inferring the operational state of electrical appliances within a home given aggregate consumption measurements collected at a single sensing point and was first introduced in the seminal paper by Hart [11]. The application of VI to NILM is not new [12, 13], however, previous approaches relied on the assumption that the system dynamics can be factorized and can be modeled by a Factorial Hidden Markov Model [14]. Note that NILM is a challenging problem because the latent variable is usually assumed to be binary, i.e. $z_t \in Z = \{0, 1\}^C$ where C is the number of components to be inferred. This integrality constraint is challenging for two reasons. First, linearizing approaches like Extended Kalman filters become hard to apply and enumerating the latent domain is computationally intractable because the cardinality of Z grows exponentially with C. As we will show later, just like VI generalizes Expectation Maximization to latent variable models with intractable posteriors, the main contribution of this paper is the generalization of VI to a class of latent variable models with intractable joint distributions. Specifically, this class constitutes dynamical systems such as those described by (1).

2. VARIATIONAL INFERENCE

We begin by deriving Variational Inference from the Expectation Maximization (EM) algorithm. The EM algorithm [15] is an algorithm to perform maximum likelihood inference on unknown parameters Θ in a latent variable model governed by observations x and latent variables z, i.e. it maximizes $\sum_z p_\Theta(x, z)$ w.r.t. Θ. It can be shown that EM performs coordinate ascent on a function F_1 known as the Variational Free Energy defined by:

$$F_1(\Theta, \tilde{P}) = \mathbb{E}_{\tilde{P}} \log p_\Theta(x, z) - \mathbb{E}_{\tilde{P}} \log \tilde{p}(z)$$

Specifically, maximizing F_1 in the direction of $\tilde{P}$ (the E-step) computes $\tilde{P} = p(z|x)$ whereas the maximization step

in the direction of Θ (the M-step) improves the evidence $p_\Theta(x)$ [16]. Therefore, because the E-step requires computing the posterior, EM is only applicable if computing the posterior $p(z|x)$ is computationally tractable. However, for many latent variable models, computing the posterior is computationally intractable because the denominator of $p(z|x) = \frac{p(z,x)}{\sum_{z \in Z} p(z,x)}$ is oftentimes hard to compute if the support of the latent variable is large. Variational Inference (VI) is a generalization of EM to latent variable models with intractable posterior distributions [17].

The main idea behind Variational Inference is the introduction of a tractable auxiliary distribution Q_ψ parameterized by the variational parameters ψ. Q_ψ is chosen from a family of distributions such that ideally, there is a ψ such that $q_\psi(z|x) = p_\Theta(z|x)$ and because of recent successes of Neural Networks for non-linear optimization, Q_ψ is often parameterized by Neural Netsworks. For VI, a function akin to F_1 is maximized, which substitutes $\tilde{P}$ for the auxiliary but tractable distribution Q_ψ.

$$\begin{aligned} F_2(\Theta, \psi) &= \mathbb{E}_{q_\psi(z|x)} \log p_\Theta(x, z) - \mathbb{E}_{q_\psi(z|x)} \log q_\psi(z|x) \\ &= \log p_\Theta(x) - D_{KL}(q_\psi(z|x)||p_\Theta(z|x)) \end{aligned}$$

Maximizing F_2 w.r.t. Θ optimizes a lower bound of the evidence. This bound is tight if $q_\psi(z|x) = p_\Theta(z|x)$, i.e. $D_{KL}(q(z|x)||p_\Theta(z|x)) = 0$. Furthermore, maximizing F_2 w.r.t. ψ minimizes the KL-divergence, i.e. it tightens the bound. Note that, because of these properties, VI also allows for performing posterior inference. After the optimization procedure, because Q_ψ will be maximally similar to P_Θ, in order to perform posterior inference on the intractable P_Θ, inference is performed on Q_ψ instead. However, note that although Variational Inference generalizes the EM algorithm to latent variable problems with intractable posterior distributions, it still requires the joint distribution $p_\Theta(x, z)$ to be tractable. However, for many latent variable models even the joint distribution might be intractable. One such class of problems constitute temporal models. In this paper, we generalize VI to this class of problems.

3. INTRACTABLE JOINT

The class of latent variable models of interest constitute dynamical systems in which the latent state evolves over time according to dynamics adhering to the first-order Markov assumption and the observation is some probabilistic function of the system state. This entails that the joint distribution of the observation and system states can be factored based on $p(x_t|z_t)$ and $p(z_t|z_{t-1})$[1], i.e.:

$$p(x_{1:T}, z_{0:T}) = p(z_0) \prod_{t=1}^{T} p(x_t|z_t) p(z_t|z_{t-1})$$

[1]For notational convenice, all dependencies on parameters Θ and ψ are omitted.

Let $p_t = p(z_t|x_{1:t})$ and $p_t = p(z_t|x_{1:t})$. For such a model, a lower bound of the chain-rule factorization of the likelihood can be derived as:

$$\begin{aligned} F_3(\Theta, \psi, t) &= \mathbb{E}_{q_t} \log p_\Theta(x_t, z_t|x_{1:t-1}) - \mathbb{E}_{q_t} \log q_t \quad (2) \\ &= \log p_\Theta(x_t|x_{1:t-1}) - D_{KL}(q_t||p_t) \quad (3) \end{aligned}$$

Note that summing F_3 over time steps implies that a lower bound of the log-evidence is maximized since:

$$\sum_{t=0}^{T} F_3(\Theta, \psi, t) = \log p_\Theta(x_{1:T}) - \sum_{t=1}^{T} D_{KL}(q_t||p_t)$$

However, evaluating the bound in 2 is intractable for many latent variable models of interest because the joint distribution is intractable as seen below:

$$p(x_t, z_t|x_{1:t-1}) = p(x_t|z_t) \sum_{z' \in Z} p(z_t|z') p(z'|x_{1:t-1}) \quad (4)$$

First, the summation over the latent domain is usually intractable. Second, evaluating equation (4) at time point t requires knowledge of the posterior at time $t-1$ which is intractable.

3.1. Monte Carlo Integration and Importance Sampling

In the following, we will show how an unbiased approximation can be obtained that makes use of Monte Carlo Integration in conjunction with Importance Sampling [18].

Monte Carlo (MC) Integration is a numerical technique to approximate an expectation of the type $\mathbb{E}_{p(z)} f(z)$ by sampling, i.e. N samples are drawn i.i.d. from $p(z)$ and $\mathbb{E}_{p(z)} f(z) \approx \frac{1}{N} \sum_{i=1}^{N} f(z^{(i)})$ with $z^{(i)} \sim p(z)$.

Note that the intractable summation in equation (4) can be written as an expectation of this type, i.e.

$$p(x_t, z_t|x_{1:t-1}) = p(x_t|z_t) \mathbb{E}_{p_{t-1}} p(z_t|z_{t-1}) \quad (5)$$

However, drawing samples from $p(z_{t-1}|x_{1:t-1})$ is not trivial and would require time-consuming advanced samplers. Instead, a technique to change the sampling distribution called Importance Sampling is being employed.

Importance Sampling is usually used as a variance reduction technique. However, in this case, it will be used to ease the computational burden of approximating equation (5). The general idea is the following: Sampling from $p(z_{t-1}|x_{1:t-1})$ is computationally challenging, however, we have access to a distribution similar to P from which sampling is, in comparison, computationally cheap: the auxiliary distribution Q. Thus, we can rewrite equation (5) in the following way:

$$p(x_t, z_t|x_{1:t-1}) = p(x_t|z_t) \mathbb{E}_{q_{t-1}} \frac{p_{t-1}}{q_{t-1}} p(z_t|z_{t-1}) \quad (6)$$

If $q_t = 0$ entails $p_t = 0$, then equation (6) is an unbiased estimator of equation (4). However, in order to evaluate equation

(6) knowledge of the true posterior $p_{t-1} = p(z_{t-1}|x_{1:t-1})$ is required which was deemed intractable.
Note that $p(z_t|x_{1:t-1}) = \frac{p(z_t,x_t|x_{1:t-1})}{p(x_t|x_{1:t-1})}$, thus if $p(x_t|x_{1:t-1})$ was provided, the joint distribution could be computed recursively. However, $p(x_t|x_{1:t-1})$ is computationally intractable because it would require enumeration of the latent space. Instead, an asymptotically unbiased estimation of $p(x_t|x_{1:t-1})$ is obtained by, again, making use of MC Integration in conjunction with Importance Sampling:

$$\hat{p}(x_t|x_{1:t-1}) = \mathbb{E}_{q_t} \frac{p(x_t, z_t|x_{1:t-1})}{q_t}$$

Putting these findings together yields what is known as self-normalizing Importance Sampling [18]. A density w is defined as follows:

$$w_{t-1} = \frac{p(z_t, x_t|x_{1:t-1})}{q(z_t|x_{1:t})} \frac{1}{\hat{p}(x_t|x_{1:t-1})} \tag{7}$$

A tractable and asymptotically unbiased approximation of equation (4) can therefore be obtained by evaluating:

$$p_\Theta(x_t, z_t|x_{1:t-1}) = p(x_t|z_t)\mathbb{E}_{q_{t-1}} w_{t-1} p(z_t|z_{t-1}) \tag{8}$$

Note that by making use of the approximation described in equation (8), equation (2) can be evaluated by Monte Carlo Integration and that F_3 can be maximized by obtaining gradient estimators with techniques introduced in [4]. However, even though the gradient estimator is unbiased, it usually has high variance making learning difficult.

4. VARIANCE REDUCTION

It is well known that Variational Inference struggles with high variance estimators and numerous techniques for variance reduction have been proposed based on e.g. Rao-Blackwellization, control variates, reparameterization [6] as well as quasi-Monte Carlo techniques [19]. Note that usually, the gradient estimator w.r.t. ψ, i.e. the gradients of the auxiliary distribution (in this case the neural network weights), are subject to high variance whereas gradients w.r.t. Θ are less problematic. In the following section two variance reduction techniques tailored to the problem at hand are introduced.

4.1. Sampling without replacement

Because the system state is assumed to be discrete, a variance reduction technique that has been studied for decades, namely *sampling without replacement* can be applied. With the correct choice of the sampling scheme, the variance of the estimator can be reduced considerably whilst not introducing a bias [20]. In addition to a reduction in variance, sampling without replacement avoids the problem of mode collapse. When sampling with replacement, the system is at danger of erroneously assigning all the probability mass to a single latent state z. If this is the case, the algorithm has essentially stopped exploring the latent domain and 'got stuck'. Note that sampling w/o replacement from Q is not trivial. However, there is considerable body of pre-existing work. We follow the scheme introduced in [21] with some slight modifications. Instead of using the Pareto design as the underlying sampler, in this work an elimination sampler introduced in [22] was employed. This results in a slower but more accurate sampling design.

4.2. Control Variate

It can be shown that the gradient estimator of F_3 w.r.t. the variational parameters ψ is an unbiased estimator of the gradient of the KL-divergence [23], i.e. [2]:

$$\underbrace{\mathbb{E}_{q(z|x)} \nabla_\psi \log \frac{p(z|x)}{q(z|x)}}_{\text{KL-divergence}} = \underbrace{\mathbb{E}_{q(z|x)} \nabla_\psi \log \frac{p(x,z)}{q(z|x)}}_{F_3}$$
$$= \underbrace{\mathbb{E}_{q(z|x)} \nabla_\psi \log \frac{p(x,z)}{q(z|x)} - c}_{\text{Control Variate}}$$

However, the variance of the gradient of F_3 exhibits much more variance. This is why control variates have been proposed. It can be shown that any constant c can be subtracted from F_3 without introducing a bias. The question then is which c to use. Note that if $c = \log p(x)$, then an estimator with the variance of the KL-divergence estimator is obtained and also note that in 3.1 a method to obtain an approximation of $p(x)$ was introduced. Using $c = \log p(x)$ is however not optimal but worked well in our experiments. Using this control variate also simplifies the implementation, since if $c_t = \log \hat{p}(x_t|x_{1:t-1})$, then:

$$F_4(\Theta, \psi, t) = \mathbb{E}_{q_\psi(z_t|x_{1:t})} \log \frac{p_\Theta(x_t, z_t|x_{1:t-1})}{\log q_\psi(z_t|x_{1:t})} - c_t$$
$$= \mathbb{E}_{q_\psi(z_t|x_{1:t})} \log w_t$$

Thus, using $c = \log p(x)$ as a control variate reduces the algorithm to recursively computing $\log w_t = \log w(z_t|x_{1:t})$ as defined in equation (7). Note that for all optimization steps, we treat c as a constant, i.e. even though c depends on Θ, we do not allow gradients to flow into c.

Below, the algorithm we call *Neural Variational Identification and Filtering* (NVIF) is described in pseudo-code. Note that the algorithm recycles samples: In order to compute $w(z^{(i)}|x_{1:t})$ samples of $w(z_{t-1}|x_{1:t-1})$ are required. In order to avoid excessive sampling, for all $i \in [1 \,..\, N]$, the same set of samples from the previous time step are used to compute $w(z^{(i)}|x_{1:t})$.

[2] Note that, for notational convenience, temporal dependencies are omitted. In other words, this is also true for F_2

Algorithm 1 Neural Variational Identification and Filtering

for $t \in [1 \,..\, T]$ **do**
  for $i \in [1 \,..\, N]$ **do**
    $z^{(i)} \sim q(z_t|x_{1:t})$ ▷ Sample w/o replacement
    Compute $p(x_t, z^{(i)}|x_{1:t-1})$ according to (8)
  end for
  Compute $\hat{p}(x_t|x_{1:t-1})$ based on all $z^{(i)}$
  Compute and store $w(z^{(i)}|x_{1:t})$ based on (7) for all i
  Gradient step to maximize F_4 w.r.t. ψ and Θ
end for

5. EXPERIMENTS

As stated earlier, experiments are conducted in the context of Non-Intrusive Load Monitoring on the REDD dataset [24]. In the following, the dynamical system model and choice of auxiliary distribution are described. Note that the goal of this paper is not to design the optimal model for appliance behavior but to showcase a novel algorithm for learning and inference in non-linear stochastic dynamical systems. However, as we will show later, even though, the model of appliance behavior is not refined, the model achieves results comparable to state of the art algorithms.

Observed Variable Like in [12], instantaneous power waveforms extracted between zero-crossings constitute x_t

Observation Because instantaneous power is an additive quantity, $p_\Theta(x_t|z_t) = \mathcal{N}(x_t; Wz_t, \sigma I)$ with W constituting unknown component waveforms.

Dynamics In order to suppress rapid switching of components, dynamics are chosen that penalize the number of components that switch: $p(z_t|z_{t-1}) = \frac{S(||z_t - z_{t-1}||)}{\sum_{j=0}^{C} \binom{C}{j} S(j)}$ with S assigning a penalty to each number of potential switches. S is not learned but kept fixed, i.e. $\Theta = \{W\}$

Aux. distribution We make use of the auxiliary distribution introduced in [8]. Additionally, because temporal dependencies are modeled, a recurrent Neural Network, in particular an LSTM, is employed.

5.1. Results

The algorithm was run for 300 epochs and 15 components ($C = 15$) were inferred. For all appliances provided as ground truth, the component with the highest mean precision-recall was chosen, just as in [25]. *NVIF* is compared to VarBOLT, another state-of-the-art VI-based model for FHMMs that is considerably less scalable and makes application-specific assumptions and NFHMM, a non-parametric framework for unsupervised learning in FHMMs. Even though VarBOLT is hand-tailored to NILM, *NVIF* shows comparable performance with $N = 500$ as shown in Table 1.

However, the more important evaluation criterion is the sample-efficiency, i.e. how many samples (N) are required to achieve results comparable to EM if it were computationally tractable. Note that because samples are drawn without replacement, if N approaches 2^C, *NVIF* becomes the EM-algorithm. Increasing N is not expected to increase performance beyond a given point and the question arises when this point is reached.

(a) Circuit	NVIF	VarBOLT
Microwave	100% / 0.1%	88.8% / 8.0%
Bath GFI	78.5% / 65.3%	71.9% / 40.2%
Electronics	90.7% / 41.2%	87.8% / 40.7%
Kitch. Out. 1	99.5% / 1.9%	8.6% / 32.8%
Furnace	66.4% / 54.2%	85.0% / 50.6%
Kitch. Out. 2 .	5.8% / 46.7%	5.3% / 70.1%
Washer/Dryer	89.5% / 64.1%	97.3% / 72.3%

(b)	NFHMM	VarBOLT	NVIF
Overall panel	0.25	0.63	0.59

Table 1. (a) Performance compared to VarBOLT as measured in Precision / Recall [25]. (b) Performance comparison with NFHMM and VarBOLT in GSPA [26].

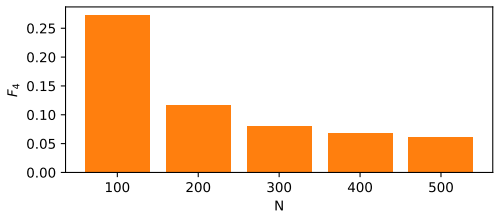

Fig. 1. Performance measured by F_4 as a function of number of samples N after convergence (300 epochs).

Figure 1 shows F_4 after convergence (300 epochs) for different numbers of samples N. One can see that the performance saturates quickly. The increase in performance from 400 to 500 is minuscule. Thus, by only exploring only about $1 - 2\%$ of the latent space, *NVIF* achieves promising results.

6. CONCLUSION

To sum up, in this paper, an asymptotically unbiased (given an appropriate choice of auxiliary distribution) algorithm for learning and inference in dynamical systems was introduced and evaluated in the context Non-Intrusive Load Monitoring. The algorithm was shown to be sample-efficient and even with a naïve model of NILM showed comparable results to existing algorithms. The introduced algorithm is general in nature and could in principle be applied to any dynamical system with binary latent variables and be generalized to continuous latent variables by replacing the auxiliary distribution.

7. REFERENCES

[1] Lennart Ljung, "System identification," in *Signal analysis and prediction*, pp. 163–173. Springer, 1998.

[2] Rudolf Emil Kalman et al., "Contributions to the theory of optimal control," *Bol. Soc. Mat. Mexicana*, vol. 5, no. 2, pp. 102–119, 1960.

[3] Peter Van Overschee and BL De Moor, *Subspace identification for linear systems: TheoryImplementationApplications*, Springer Science & Business Media, 2012.

[4] Rajesh Ranganath, Sean Gerrish, and David Blei, "Black box variational inference," in *Artificial Intelligence and Statistics*, 2014, pp. 814–822.

[5] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[6] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] Danilo Jimenez Rezende and Shakir Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

[8] Henning Lange and Mario Bergés, "Factornet: Learning to factorize intractable and multi-modal posterior distributions for energy disaggregation," in *Proceedings of the 4th International Workshop on Non-Intrusive Load Monitoring*, 2018.

[9] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt, "Advances in variational inference," *arXiv preprint arXiv:1711.05597*, 2017.

[10] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[11] George William Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[12] Henning Lange and Mario Bergés, "Variational bolt: Approximate learning in factorial hidden markov models with application to energy disaggregation," in *Proceedings of the AAAI*, 2018.

[13] Yin Cheng Ng, Pawel M Chilinski, and Ricardo Silva, "Scaling factorial hidden markov models: stochastic variational inference without messages," in *Advances in Neural Information Processing Systems*, 2016, pp. 4044–4052.

[14] Zoubin Ghahramani and Michael I Jordan, "Factorial hidden markov models," in *Advances in Neural Information Processing Systems*, 1996, pp. 472–478.

[15] 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 (methodological)*, pp. 1–38, 1977.

[16] Radford M Neal and Geoffrey E Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, pp. 355–368. Springer, 1998.

[17] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[18] Art B. Owen, *Monte Carlo theory, methods and examples*, 2013.

[19] Alexander Buchholz, Florian Wenzel, and Stephan Mandt, "Quasi-monte carlo variational inference," *arXiv preprint arXiv:1807.01604*, 2018.

[20] Daniel G Horvitz and Donovan J Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[21] Rohan Shah and Dirk P Kroese, "Without-replacement sampling for particle methods on finite state spaces," *Statistics and Computing*, vol. 28, no. 3, pp. 633–652, 2018.

[22] Jean-Claude Deville and Yves Tille, "Unequal probability sampling without replacement through a splitting method," *Biometrika*, vol. 85, no. 1, pp. 89–101, 1998.

[23] Andriy Mnih and Karol Gregor, "Neural variational inference and learning in belief networks," *arXiv preprint arXiv:1402.0030*, 2014.

[24] J Zico Kolter and Matthew J Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, 2011, vol. 25, pp. 59–62.

[25] J Zico Kolter and Tommi Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," in *Artificial Intelligence and Statistics*, 2012, pp. 1472–1482.

[26] Ruoxi Jia, Yang Gao, and Costas J Spanos, "A fully unsupervised non-intrusive load monitoring framework," in *Smart Grid Communications (SmartGridComm), 2015 IEEE International Conference on*. IEEE, 2015, pp. 872–878.