Structure Inference in Sum-Product Networks using Infinite Sum-Product Trees

Martin Trapp[1,2], **Robert Peharz**[3], **Marcin Skowron**[1], **Tamas Madl**[1], **Franz Pernkopf**[2], **and Robert Trappl**[1]

[1]Austrian Research Institute for Artificial Intelligence
[2]Graz University of Technology
[3]Medical University of Graz

Abstract

Sum-Product Networks (SPNs) are a highly efficient type of a deep probabilistic model that allows exact inference in time linear in the size of the network. In previous work, several heuristic structure learning approaches for SPNs have been developed, which are prone to overfitting compared to a purely Bayesian model. In this work, we propose a principled approach to structure learning in SPNs by introducing infinite Sum-Product Trees (SPTs). Our approach is the first correct and successful extension of SPNs to a Bayesian nonparametric model. We show that infinite SPTs can be used successfully to discover SPN structures and outperform infinite Gaussian mixture models in the task of density estimation.

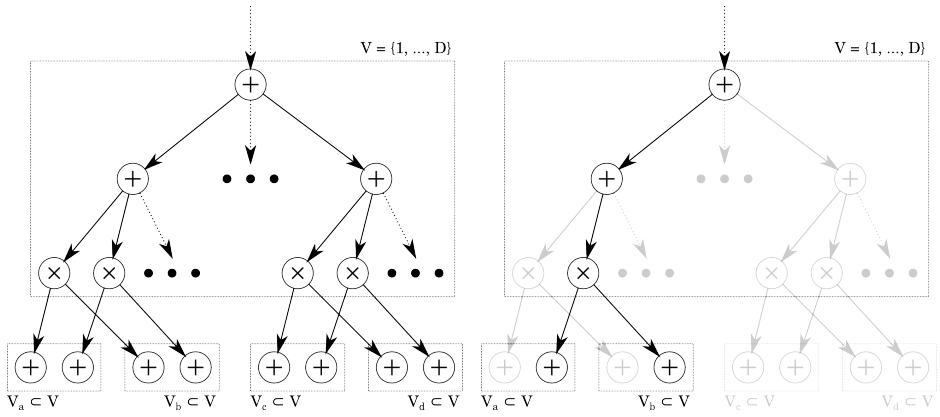
1 Introduction

Sum-Product Networks (SPNs) [6, 21] are a highly efficient type of deep probabilistic models and have been successfully applied to various generative and discriminative tasks, e.g. [21, 12, 5, 18, 2]. In recent years, several approaches to structure and parameter learning of SPNs have been proposed, e.g. [8, 13, 18, 20, 24, 1, 9, 28, 29]. SPNs can be defined recursively, as weighted sums and products of smaller SPNs, with univariate or multivariate probability distributions as leaf nodes. In a complete and decomposable SPN [21, 17], all children of a sum node have the same variable scope as the sum, whereas the children of each product partition the product's scope into non-empty disjoint sub-scopes. Complete and decomposable SPNs can be represented as a sum of induced trees ([28, 29]). The generative process of normalized SPNs can be described informally by: (1) selecting an induced tree $\mathcal{T}$ with probability proportional to its weights: $P(\mathcal{T}) \propto \prod_{w \in \mathcal{T}} w$ and (2) sampling the observation from the leaf node distributions of the induced tree. Therefore, the posterior distribution of any Bayesian nonparametric extension of SPNs depends on the notion of induced trees. We introduce infinite Sum-Product Trees (SPTs), the first Bayesian nonparametric extension of SPNs with a posterior distribution based on induced trees. Previous work ([15]) neglected induced trees in their posterior construction and did not report quantitative results and comparisons to existing approaches. We show that our infinite SPTs allow to discover SPN structures with high modelling performance while maintaining a good generalisation behavior.

2 Model

In order to define our generative process for infinite SPTs, the following simplifying but not very restrictive assumptions are made: (1) All leaf distributions are univariate; (2) all product nodes have two children; (3) sums and products occur in an alternating fashion and (4) the root node is a sum. Additional to those assumptions, we augment an SPN structure with so-called group nodes. Each

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.



(a) Building block of infinite SPTs. (b) Example of an induced tree in infinite SPTs.

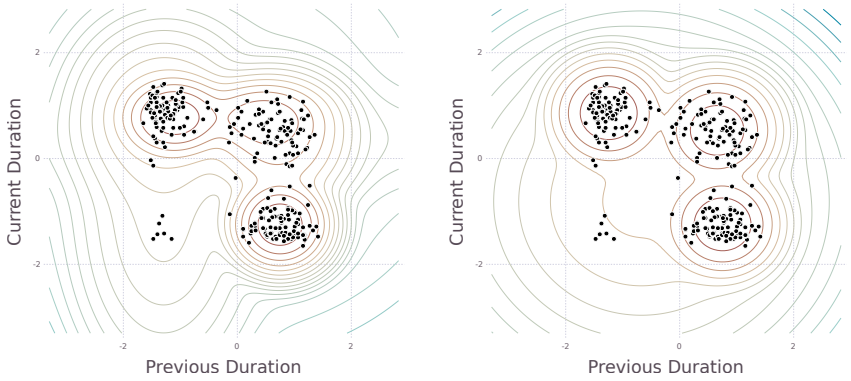
Figure 1: In infinite SPTs, sums of a SPN are augmented by additional sums (group nodes) one level below them, which allow product nodes that apply the same partition of their scope to be grouped together. Note that $V_i \neq V_j \; \forall (i,j) \in \{a,b,c,d\}$ and all V_i with $i \in \{a,b,c,d\}$ are non-empty sub-scopes of V. Moreover, $V_a \cup V_b = V$, $V_a \cap V_b = \emptyset$ and $V_c \cup V_d = V$, $V_c \cap V_d = \emptyset$.

group node, which is essentially an additional sum, groups together products which share the same partition of their scope. Figure 1 illustrates the basic structure of infinite SPTs with additional group nodes and induced trees in the model construction. Note that the number of group nodes, children of sum S with variable scope V, is equal to $\left\{ {|V| \atop 2} \right\}$, where $\left\{ \right\}$ is the Stirling number of the second kind and $|V|$ is the cardinality of the scope. In order to enumerate each partition of scope $V = \{1, \ldots, D\}$, we index every possible partition of scope into two non-empty disjoint sub-scopes resulting in the index set $U = \{1, \ldots, \left\{ {|V| \atop 2} \right\}\}$. Similar to the definition of SPNs, the generative process of infinite SPTs is defined recursively. Consider N observations with dimensionality D, where we denote x_n^d to be the value of the d^{th} variable of observation $n \leq N$. The generative process of infinite SPTs can be described as follows:

Starting at the root node:

1. If the scope $V_S \subseteq V$ for the current node S is multivariate:
   - Draw weights $\boldsymbol{w}_S$ to the $\left\{ {|V_S| \atop 2} \right\}$ group nodes from a Dirichlet distribution with hyper-parameter α_S.
   - According to $\boldsymbol{w}_S$ draw the latent assignments $c_{S,n}$ to the group nodes and draw the partition assignment of each group node $u_{c_{S,n}} \in U$ uniformly without replacement.
   - For each selected group node, draw latent assignments $z_{c_{S,n}}$ for the observations at the group node from a Chinese restaurant process with hyper-parameter $\beta_{c_{S,n}}$ to product nodes.
   - For each selected product, partition the scope into non-empty disjoint sub-scopes according to $u_{c_{S,n}}$ and for each child of the product, apply the infinite SPN process for the observations at the child recursively.
2. Else for node S with univariate scope $d \in V$:
   - Draw latent leaf assignments $c_{S,n}$ from a Chinese restaurant process with hyper-parameter γ_S and draw distribution parameters $\theta_{c_{S,n}}$ from an appropriate prior.
   - Generate the value of the d^{th} dimension for the n^{th} observation from an appropriate leaf node distribution parametrized with $\theta_{c_{S,n}}$.

Our MCMC inference algorithm uses Gibbs sampling and is based on the work by [16]. The algorithm scales linearly in the number of observations and active sums in the network and includes inference over the hyper-parameters. Detailed descriptions of the MCMC inference algorithm are omitted due to space constraints. With increasing dimensionality of the data, the infinite SPT learns increasingly complex and deep hierarchical representations of the data. Similar to the crosscat model [26] the



(a) Log density of an infinite SPT. (b) Log density of an infinite GMM.

Figure 2: Visualisation of Old Faithful geyser with the corresponding log density modelled by an infinite SPT and an infininte GMM at iteration 10^3. To allow for better comparability, both models are learned using the same hyperparameter settings and base distributions.

Table 1: Average 10-fold cross-validation log predictive densities and Mann-Whitney U p-values. The infinite GMM and the infinite SPT are trained over 1000 iterations with Gibbs sampling.

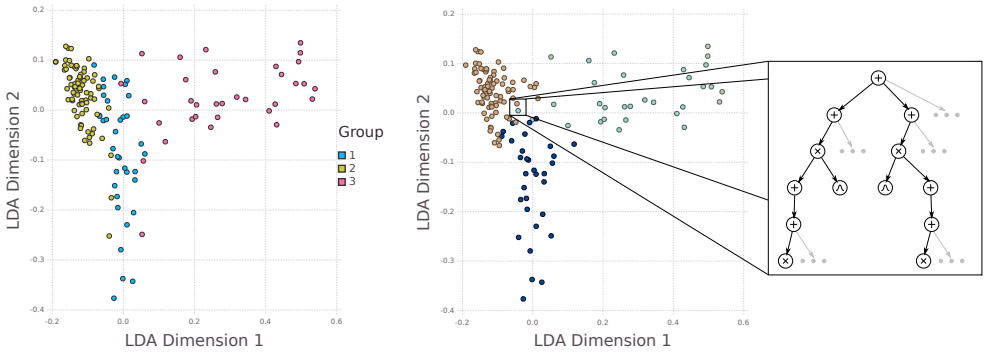
Dataset	infininte GMM	infinite SPT	infinite GMM / infinite SPT
Old Faithful	-1.737	-1.700	< 0.01
Chemical Diabetes	-3.022	-2.879	< 0.01
Iris	-3.943	-3.744	< 0.01

infinite SPT has asymptotic capacity in terms of the number of observations and the dimensionality of the data. Moreover, the infinite SPT can be used to analyse heterogeneous data. In contrast to previous work in the field, e.g. [3, 14, 26], observations are generated from induced trees, which are selective SPNs ([19]) inside the infinite SPT.

3 Experiments and Future Work

We present results on artificial and real-world datasets with different dimensions and complexity. To assess the performance of our approach we measure the density estimation performance of the infinite SPT on the Old Faithful geyser dataset (processed as described by [25]), the Chemical Diabetes dataset [23], (using the dimensions: glucose area, insulin area and insulin resistance) and the Iris dataset [11]. In addition, we demonstrate latent variable discovery capacities of infinite SPTs on the Chemical Diabetes and a movie script dialogs dataset. In all experiments, we initialize our infinite SPT using a sequential construction of the structure. We used conjugate priors and sampled the hyper-parameters for the Chinese restaurant processes from independent Gamma priors [10].

Density Estimation Figure 2 illustrates that infinite SPTs are better able to explain the data compared to infinite Gaussian mixture models (GMM) [22]. Infinite SPT favor deep structures, due to the construction of the prior on induced trees, allowing the model to fit complex distribution more easily than a shallow model ([7]), e.g. infinite GMMs. With growing dimensionality, the distribution modelled by an infinite SPT will differ from those modelled by an infinite GMM with increasing probability. In our experiments we used a univariate Normal-Normal distribution as base distribution for all methods. Table 1 shows the average 10-fold cross-validation log predictive densities computed as in [27] for (1) infinite GMMs and (2) infinite SPTs. On all datasets, the infinite SPT obtains significantly better average 10-fold cross-validation log predictive densities than infinite GMMs.



(a) True group assignments.

(b) Projected group assignments.

Figure 3: Comparison of true group assignments and group assignments estimated by an infinite SPT on the Chemical Diabetes dataset. The coloring of the projected assignments encodes the position of their induced trees on a one dimensional embedding of the differences between assignment vectors, e.g. dark blue dots are more similar to light blue dots then to red dots. The enlarged section illustrates the induced trees of two observations from the dataset.

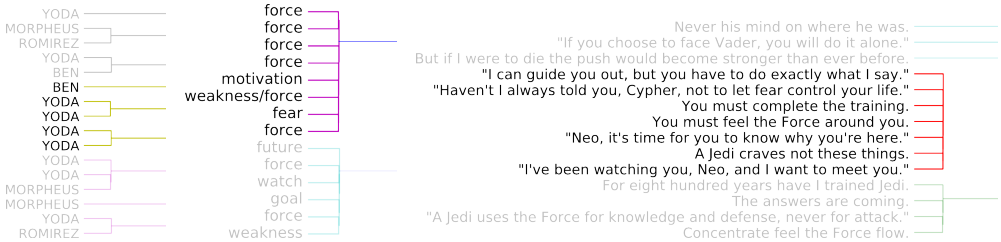


Figure 4: Preliminary results on the movie script dialog dataset. The infinite SPT finds groupings of the utterances based on their one-hot coding related to similar movie source, similar latent characteristic and similar dialog act, from left to right.

Latent Variable Discovery In addition, we analysed the clustering capacities qualitatively on the Chemical Diabetes dataset and show preliminary results on a complex real-world dataset of movie script dialogs. The resulting grouping on the Chemical Diabetes dataset is illustrated in Figure 3 with positions obtained from a linear discriminant analysis. In order to visualize the group assignments, we used multidimensional scaling [4] to transform the binary assignment vectors (each representing an induced tree inside the infinite SPT) into an one dimensional space. Alternatively, the binary assignment vectors could be clustered using a hierarchical clustering, allowing for inspection of the hierarchy learned by the infinite SPT. The real-world dataset of movie script dialogs is based on utterances of four different mentor characters: Yoda and Obi-Wan Kenobi from Star Wars, Morpheus from The Matrix and Juan Sánchez Ramírez from Highlander. We extracted 62 utterances and used a one-hot coding based on the 25 most frequent words (excluding stop words) to encode each utterance. Moreover, each utterance was manually annotated by a human expert in terms of a latent characteristic, e.g. weakness revealing utterance. We use an infinite SPT model with Beta-Bernoulli distributions to identify latent groupings of those utterances. Preliminary results are shown in Figure 4.

Future Work In future work we will explore: (1) implementation of infinite SPT exploiting parallel computations; (2) hierarchical priors for leaf node parameters and (3) explore alternative representations that allow for more flexible and complex network structures than the infinite SPT, e.g. node sharing. As the current experiments show promising results, we will conduct a wider variety of experiments to investigate the capabilities of our approach in more detail.

Acknowledgments

This research is partially funded by the Austrian Science Fund (FWF) under grant no. P 27530.

References

[1] T. Adel, D. Balduzzi, and A. Ghodsi. Learning the structure of sum-product networks via an SVD-based algorithm. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 32–41, 2015.

[2] M.R. Amer and S. Todorovic. Sum-product networks for modeling activities with stochastic structure. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1314–1321, 2012.

[3] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):2–30, 2010.

[4] I. Borg and P. J. F. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[5] W.C. Cheng, S. Kok, H.V. Pham, H.L. Chieu, and K.M.A. Chai. Language modeling with sum-product networks. In *Annual Conference of the International Speech Communication Association (Interspeech)*, 2014.

[6] A. Darwiche. A differential approach to inference in Bayesian networks. *Journal of the ACM (JACM)*, 50(3):280–305, 2003.

[7] O. Delalleau and Y. Bengio. Shallow vs. deep sum-product networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[8] A. Dennis and D. Ventura. Learning the architecture of sum-product networks using clustering on variables. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2042–2050, 2012.

[9] A. Dennis and D. Ventura. Greedy structure search for sum-product networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 932–938, 2015.

[10] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[11] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[12] R. Gens and P. Domingos. Discriminative learning of sum-product networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3248–3256, 2012.

[13] Robert Gens and Pedro Domingos. Learning the structure of sum-product networks. *International Conference on Machine Learning (ICML)*, pages 873–880, 2013.

[14] Z. Ghahramani, M. I. Jordan, and R. P. Adams. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 19–27, 2010.

[15] S.W. Lee, C.J. Watkins, and B.T. Zhang. Non-parametric bayesian sum-product networks. In *Workshop on Learning Tractable Probabilistic Models (LTPM)*, 2014.

[16] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[17] R. Peharz. *Foundations of Sum-Product Networks for Probabilistic Modeling*. PhD thesis, Graz University of Technology, 2015.

[18] R. Peharz, B. Geiger, and F. Pernkopf. Greedy part-wise learning of sum-product networks. In *European Conference on Machine Learning (ECML)*, volume 8189, pages 612–627. Springer Berlin, 2013.

[19] R. Peharz, R. Gens, and P. Domingos. Learning selective sum-product networks. In *Workshop on Learning Tractable Probabilistic Models (LTPM)*, 2014.

[20] R. Peharz, R. Gens, F. Pernkopf, and P. Domingos. On the latent variable interpretation in sum-product networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. http://arxiv.org/abs/1601.06180.

[21] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 337–346, 2011.

[22] C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 554–560, 1999.

[23] G. M. Reaven and R. G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16(1):17–24, 1979.

[24] A. Rooshenas and D. Lowd. Learning sum-product networks with direct and indirect variable interactions. In *International Conference on Machine Learning (ICML)*, pages 710–718, 2014.

[25] D. W. Scott. Multivariate density estimation and visualization. In *Handbook of Computational Statistics*, pages 549–569. Springer, 2012.

[26] P. Shafto, C. Kemp, V. Mansinghka, and J.B. Tenenbaum. A probabilistic model of cross-categorization. *Cognition*, 120(1):1–25, 2011.

[27] A. Vehtari and J. Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002.

[28] H. Zhao, T. Adel, G. Gordon, and B. Amos. Collapsed variational inference for sum-product networks. In *International Conference on Machine Learning (ICML)*, pages 1310–1318, 2016.

[29] H. Zhao, P. Poupart, and G. Gordon. A unified approach for learning the parameters of sum-product networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.