ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

A Counterexample to Ockham's Razor and the Curse of Dimensionality: Marginalising Complexity and Dimensionality for GMMs

Benoît Frénay

University of Namur - NaDI - Faculty of Computer Science - PReCISE - HuMaLearn
rue Grandgagnage 21, B-5000 Namur - Belgium

Abstract. Ockham's razor and the curse of dimensionality are two founding principles in machine learning. First, simple models should be preferred to complex ones, in order to prevent overfitting. Second, high-dimensional spaces should be avoided, whenever possible, because learning is easier in lower-dimensional spaces. These principles are often invoked to justify methodological choices or to preprocess data. However, this paper shows a counterexample where it is better to first learn a more complex model in a higher-dimensional space, and then to go back to the lower-dimensional space while dropping the additional complexity. Specifically, experiments demonstrate that Gaussian mixtures models can be learned in a higher-dimensional space and then marginalised to the target dimensionality to improve probability density estimation performances. The chosen problem is deliberately simple to facilitate the analysis, but it opens the way to similar work for more complex models and tasks.

1 Introduction

In machine learning, a common belief is that one should avoid to use complex models (Ockham's razor) and to work in high-dimensional spaces (curse of dimensionality [1, 2, 3]) as much as possible, to achieve better generalisation. This paper argues that this is not necessarily true and proposes a methodology to take advantage of high-dimensional data as an intermediary step.

Given a high-dimensional (HD) dataset $\mathcal{D}^{\text{HD}} = \left\{\mathbf{x}_1^{\text{HD}}, \ldots, \mathbf{x}_n^{\text{HD}}\right\}$ collected with $d > 1$ features $X_1, \ldots, X_d$, let us assume that one would like to obtain a model based on a subset of p of these features. For example, the dataset has been collected with some features that are no longer available, the number of considered features should remain low due to interpretablity requirements, one is interested only in univariate analysis for domain-specific reasons, etc.

The common approach would be to directly learn a model from a low-dimensional (LD) version of the dataset $\mathcal{D}^{\text{LD}} = \left\{\mathbf{x}_1^{\text{LD}}, \ldots, \mathbf{x}_n^{\text{LD}}\right\}$ with only p features. The $d - p$ remaining features are simply discarded. However, this paper shows that they can be useful, *in particular when the dataset is small*. It may be better in some cases to learn a complex model in a higher dimensional space and then to go back to the lower dimensional space. The chosen problem is deliberately simple to facilitate the analysis, but it opens the way to similar work for more complex models and tasks. The main contribution is a new learning methodology based on the idea of consistency across dimensionalities and that intuitively goes against Ockham's razor and the curse of dimensionality.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

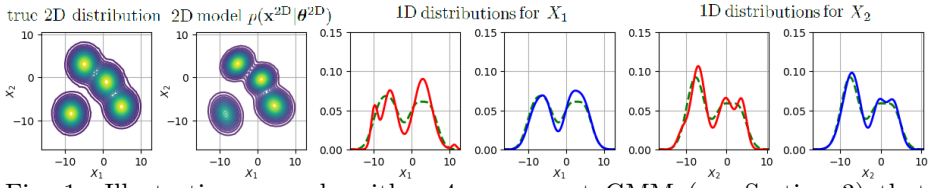

Fig. 1: Illustrative example with a 4-component GMM (see Section 3) that generated 200 training instances with features X_1 and X_2. The true 2D distribution is shown as a contour plot for X_1 and X_2 (1st plot), as well as the 2D model $p(\mathbf{x}^{2D}|\boldsymbol{\theta}^{2D})$ (2nd plot). Then, the red 1D model $p(\mathbf{x}^{1D}|\boldsymbol{\theta}^{1D})$ and the blue marginalisation $p(\mathbf{x}^{1D}|\boldsymbol{\theta}^{2D\downarrow 1D})$ of $p(\mathbf{x}^{2D}|\boldsymbol{\theta}^{2D})$ are compared to the true green 1D model $p(\mathbf{x}^{1D}|\boldsymbol{\theta}^{1D})$ for X_1 (3rd and 4th plots) and X_2 (5th and 6th plots).

The rest of this works is organised as follows. Section 2 introduces the proposed methodology based on marginalisation, which it then instantiated to Gaussian mixture models (GMMs) in Section 3. Section 4 describes the experiments performed to show the added value of the proposed methodology and discusses their results. Finally, Section 5 concludes and discusses future work.

2 Marginalisation for Consistency across Dimensionalities

This section proposes a new probabilistic methodology to learn LD models by first learning HD models as an intermediate step. The idea can be adapted to more general problems than density estimation, which is considered here.

2.1 Consistency with Respect to Marginalisation

The above problem can be tackled with a probabilistic generative model $p(\mathbf{x}|\boldsymbol{\theta})$. This model can be learned with parameters $\boldsymbol{\theta}^{\mathrm{HD}}$ and $\boldsymbol{\theta}^{\mathrm{LD}}$ in the HD or LD space with d or p features, respectively. If enough data are available, one could expect both HD and LD models to be consistent with respect to marginalisation, i.e.,

$$p(\mathbf{x}^{\mathrm{LD}}|\boldsymbol{\theta}^{\mathrm{LD}}) = \int_{\mathbf{x}^{\mathrm{HD}\setminus\mathrm{LD}}} p(\mathbf{x}^{\mathrm{HD}}|\boldsymbol{\theta}^{\mathrm{HD}}) d\mathbf{x}^{\mathrm{HD}\setminus\mathrm{LD}}, \tag{1}$$

where $\mathbf{x}^{\mathrm{HD}\setminus\mathrm{LD}}$ are the $d-p$ features that are discarded. However, in practice, this is not the case, as shown in Figure 1 where the models learned from a dataset of 200 instances in 2D and 1D are inconsistent with respect to marginalisation. In fact, if one denotes as $\boldsymbol{\theta}^{\mathrm{HD}\downarrow\mathrm{LD}}$ the parameters of the HD model that is reduced to LD

$$p(\mathbf{x}^{\mathrm{LD}}|\boldsymbol{\theta}^{\mathrm{HD}\downarrow\mathrm{LD}}) \stackrel{\mathrm{def}}{=} \int_{\mathbf{x}^{\mathrm{HD}\setminus\mathrm{LD}}} p(\mathbf{x}^{\mathrm{HD}}|\boldsymbol{\theta}^{\mathrm{HD}}) d\mathbf{x}^{\mathrm{HD}\setminus\mathrm{LD}}, \tag{2}$$

under the hypothesis that the marginalisation of $p(\mathbf{x}|\boldsymbol{\theta})$ belongs to the same model family, Figure 1 shows that $p(\mathbf{x}^{\mathrm{LD}}|\boldsymbol{\theta}^{\mathrm{HD}\downarrow\mathrm{LD}})$ is closer to the true, unknown distribution $p(\mathbf{x}^{\mathrm{LD}})$ than $p(\mathbf{x}^{\mathrm{LD}}|\boldsymbol{\theta}^{\mathrm{LD}})$. In other words, in this synthetic example, it is better to learn a more complex model in a high dimensional space first and then to marginalise it to the lower dimensional space. This is due to the fact that the modes of the distribution are easier to distinguish in two dimensions.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2.2 The LiHDaM Methodology

Motivated by the illustrative example in Figure 1, the following methodology is proposed. Given a dataset $\mathcal{D}^{\text{HD}}$ with d dimensions and a task that requires to learn a probabilistic model $p(\mathbf{x}^{\text{LD}}|\boldsymbol{\theta}^{\text{LD}})$ based on $p < d$ dimensions, one should

1. learn a model $p(\mathbf{x}^{\text{HD}}|\boldsymbol{\theta}^{\text{HD}})$ in the d-dimensional space;
2. marginalise the model to reduce it to p dimensions;
3. use the new model $p(\mathbf{x}^{\text{LD}}|\boldsymbol{\theta}^{\text{HD}\downarrow\text{LD}})$ (and drop the d-dimensional one).

The rest of this paper instantiates this learn-in-high-dimension-and-marginalise (LiHDaM) methodology to the particular case of Gaussian mixture models.

2.3 Related Work

The idea of working in a high-dimensional space is also exploited by support vector machines (SVMs) [4] that rely on the kernel trick [5]. They build an intermediate feature space whose dimensionality may even be infinite (e.g., with the RBF kernel). The advantage is that classification becomes easier in the high-dimensional feature space. However, even for SVMs, regularisation is necessary to prevent overfitting and the model complexity is therefore restricted. Here, the goal is explore whether the learning problem can be solved with a complex model which is simplified afterwards by the marginalisation mechanism. The final model lies in the LD space and all HD components are discarded. Experiments in Section 4 show that it improves generalisation and metaparameter selection.

3 Illustrative Case: Gaussian mixture models

Gaussian mixture models (GMMs) are widely-used semi-parametric models for density estimation. In hidden Markov models, they are often used to compute emission probabilities when observations are too complex for simple Gaussian distributions, like electrocardiogram (ECG) signals [6, 7]. The expression of the probability density function (PDF) for a GMM with K components is

$$p\left(\mathbf{x}\middle|\,\{\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^{K}\right)=\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{3}$$

where each component has a weight π_k, a mean $\boldsymbol{\mu}_k$ and a covariance matrix $\boldsymbol{\Sigma}_k$. Unfortunately, there exists no close-form solution to learn GMM parameters, but the iterative expectation-maximisation (EM) algorithm [8] can be used. Until convergence, the E step computes the membership γ_{ik} of $\mathbf{x}_i$ to the kth component, then the M step estimates the parameters of each component as

$$\pi_k=\frac{1}{n}\sum_{i=1}^{n}\gamma_{ik}\quad\boldsymbol{\mu}_k=\frac{\sum_{i=1}^{n}\gamma_{ik}\mathbf{x}_i}{\sum_{i=1}^{n}\gamma_{ik}}\quad\boldsymbol{\Sigma}_k=\frac{\sum_{i=1}^{n}\gamma_{ik}\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)^T\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right)}{\sum_{i=1}^{n}\gamma_{ik}}. \tag{4}$$

An interesting property of a GMM is that it can easily be marginalised: the result is just another GMM with less dimensions and parameters. Unnecessary components of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ just need to be dropped, while π_k remains unaffected. Interestingly, GMMs are easy to learn, to use and to interpret. For these reason, this paper uses GMMs to assess the interest of the LiHDaM methodology.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

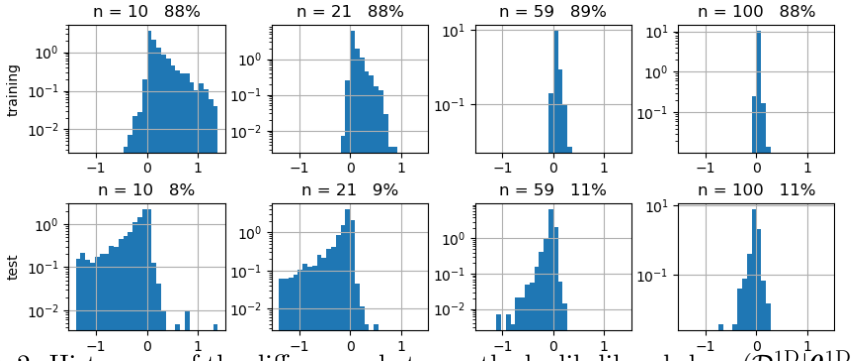

Fig. 2: Histogram of the differences between the loglikelihoods $\log p(\mathcal{D}^{1\mathrm{D}}|\boldsymbol{\theta}^{1\mathrm{D}}) - \log p(\mathcal{D}^{1\mathrm{D}}|\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}})$ computed on training (top) and test (bottom) samples, for all the GMMs trained with n instances. The percentage shown is the proportion of GMMs for which $\boldsymbol{\theta}^{1\mathrm{D}}$ is better, i.e., $\log p(\mathcal{D}^{1\mathrm{D}}|\boldsymbol{\theta}^{1\mathrm{D}}) > \log p(\mathcal{D}^{1\mathrm{D}}|\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}})$.

4 Validation of the LiHDaM Methodology

Experiments below assess the LiHDaM methodology and illustrate its key idea in a visual way with a simple 2D problem. Future works will use real-world, higher-dimensional data to further validate the benefits of model marginalisation.

4.1 Experimental Setup

For 100 repetitions, 2D isotropic GMMs with $C \in \{2, 4, 8\}$ components were created with random centers in $[10, 10] \times [10, 10]$ and width $\sigma = 2$. From each GMM, (i) several training datasets with $n \in \{10 \dots 100\} \times C$ instances were sampled, as well as (ii) a validation set half the size to select the best model complexity and (iii) a test set of 10,000 instances to assess generalisation.

The above datasets are each declined in $\mathcal{D}^{2\mathrm{D}}$ (X_1 and X_2) and $\mathcal{D}^{1\mathrm{D}}$ (X_1 only) versions. The task is to train a 1D density estimation model for X_1. GMMs were trained with $K \in \{1 \dots 10\}$ components, full covariance matrices and 100 parameter initialisations with k-means++ to obtain (i) $p(\mathbf{x}^{1\mathrm{D}}|\boldsymbol{\theta}^{1\mathrm{D}})$ from $\mathcal{D}^{1\mathrm{D}}$ as usual and (ii) $p(\mathbf{x}^{1\mathrm{D}}|\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}})$ from $\mathcal{D}^{2\mathrm{D}}$ following the LiHDaM methodology.

4.2 Experimental Results

First, Figure 2 shows that the standard model $\boldsymbol{\theta}^{1\mathrm{D}}$ is almost always ($\pm 90\%$) better on 1D training data, which is unsurprising since $\boldsymbol{\theta}^{1\mathrm{D}}$ is directly optimised on them. The marginalised model $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$ is instead obtained from 2D training data and is almost never ($\pm 2\%$) better on 1D training data. Such rare occurrences are likely due to convergence issues, as the loglikelihood is non-convex for GMMs. On the contrary, and more interestingly, the marginalised model $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$ is most often ($\pm 90\%$) equal or better on 1D test data, i.e., it offers equal or better generalisation. This is especially true when the number of instances n is small. When it is not the case, the difference in favour of $\boldsymbol{\theta}^{1\mathrm{D}}$ remains small.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

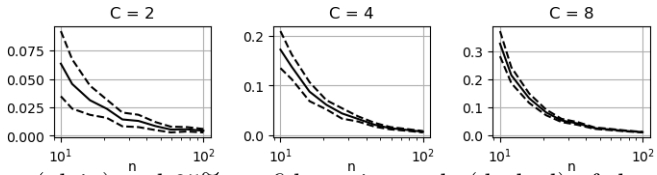

Fig. 3: Mean (plain) and 95% confidence intervals (dashed) of the difference in test loglikelihood for $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$ and $\boldsymbol{\theta}^{1\mathrm{D}}$, with different training set sizes. For each plot, the number of GMM components K is set to the number of centers C.

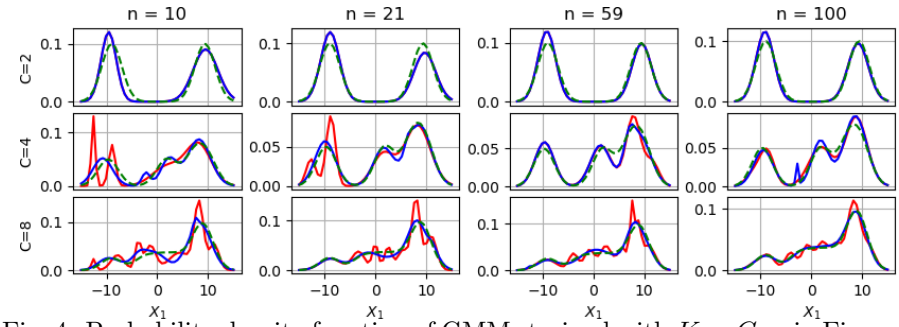

Fig. 4: Probability density function of GMMs trained with $K = C$ as in Figure 3, with different training set sizes. The green dashed line is the true distribution, whereas the red and blue plain lines correspond to $\boldsymbol{\theta}^{1\mathrm{D}}$ and $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$, respectively.

Second, Figure 3 shows that the above conclusions are even stronger when the number of components in the trained GMMs is equal to the number of centers that generated data. In that case, LiHDaM is significantly better than the traditional approach, in particular for small datasets. Figure 4 shows that the marginalised model $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$ is less likely to overfit than $\boldsymbol{\theta}^{1\mathrm{D}}$, suggesting that LiHDaM has some kind of regularisation effect, in particular for small datasets.

Third, Figure 5 shows that similar results are obtained when the number of components K is chosen so as to maximise the loglikelihood of the validation set (in 1D for $\boldsymbol{\theta}^{1\mathrm{D}}$ and in 2D for $\boldsymbol{\theta}^{2\mathrm{D}}$ that is marginalised to $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$). However, the difference is smaller than in Figure 3. Figure 5 also shows that the selected value for K is more relevant with LiHDaM, i.e., closer to the true number of components C. The marginalised model $\boldsymbol{\theta}^{2\mathrm{D}\downarrow 1\mathrm{D}}$ can afford more complexity with the same amount of training data, with less overfitting (see Figure 6).

5 Conclusion and Future Works

This paper is a first step towards better understanding when and why it may be interesting to use higher dimensional and more complex models as an intermediate step. The LiHDaM approach relies on marginalisation and could be extended to generative classification models. However, experiments will first be extended in future work to assess whether LiHDaM works for complex, real-world datasets.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

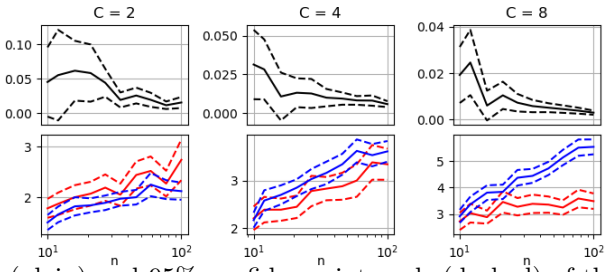

Fig. 5: Mean (plain) and 95% confidence intervals (dashed) of the difference in test loglikelihood (black, top) and of the number of components K (bottom) for $\boldsymbol{\theta}^{\mathrm{2D}\downarrow\mathrm{1D}}$ (blue) and $\boldsymbol{\theta}^{\mathrm{1D}}$ (red) when K is chosen with a validation set.

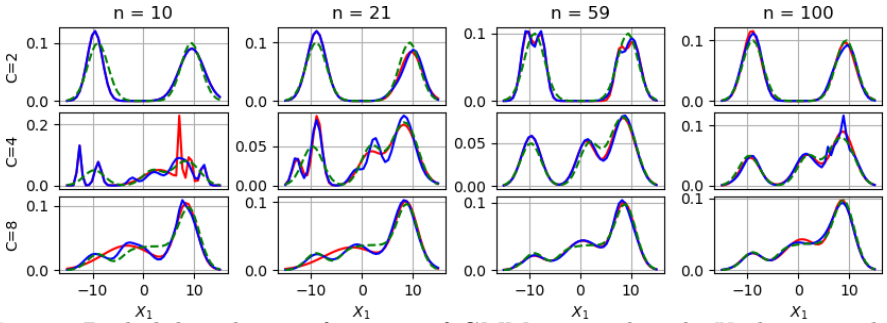

Fig. 6: Probability density function of GMMs trained with K chosen with a validation set. The green dashed line is the true distribution, whereas the red and blue plain lines correspond to $\boldsymbol{\theta}^{\mathrm{1D}}$ and $\boldsymbol{\theta}^{\mathrm{2D}\downarrow\mathrm{1D}}$, respectively.

Acknowledgments

This work is supported by the F.R.S.-FNRS EOS VeriLearn project n. 30992574 and the ARIAC by Digital Wallonia4.AI project n. 2010235 of the SPW Recherche Wallonie. The author also thanks HuMaLearn colleagues for fruitful discussions.

References

[1] Richard Bellman. Dynamic programming. *Press Princeton, New Jersey*, 1957.

[2] Mario Köppen. The curse of dimensionality. In *Proc. WSC5*, volume 1, pages 4–8, 2000.

[3] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *Proc. IWANN 2005*, pages 758–770, 2005.

[4] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[5] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[6] Gari D. Clifford, Francisco Azuaje, Patrick McSharry, et al. *Advanced methods and tools for ECG data analysis*, volume 10. Artech house Boston, 2006.

[7] Gäel de Lannoy, Benoît Frénay, Michel Verleysen, and Jean Delbeke. Supervised ECG delineation using the wavelet transform and hidden Markov models. In *Proc. ECIFMBE*, pages 22–25, 2009.

[8] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Meth*, 39(1):1–22, 1977.