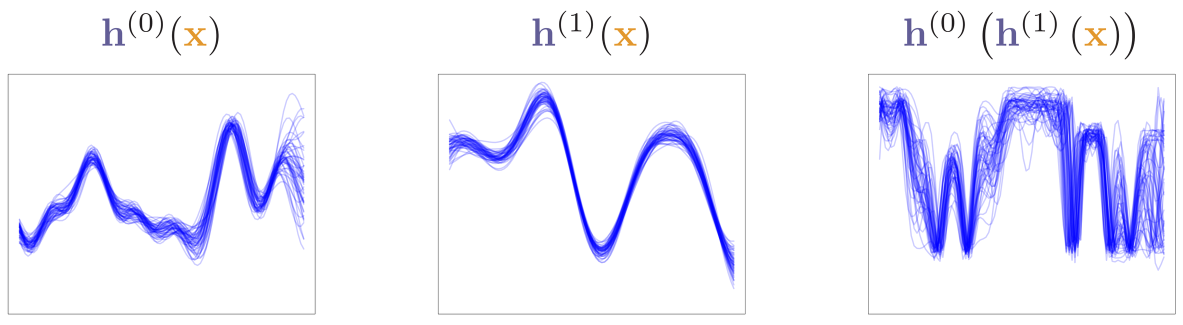
Deep Gaussian Process Autoencoders for Novelty Detection

Rémi Domingues[1], Pietro Michiardi[1], Jihane Zouaoui[2], Maurizio Filippone[1]

1 - EURECOM, Sophia Antipolis, France   2 - Amadeus, Sophia Antipolis, France

Deep Gaussian Process Autoencoders

- Unsupervised deep probabilistic model
- Suitable for any type of data, e.g. continuous, discrete, categorical
- Training only requires tensor products, no matrix factorization
- Inference through mini-batch based stochastic variational inference
- Composition of functions: $\mathbf{f}(\mathbf{x}) = \left(\mathbf{h}^{(N_h-1)}\left(\theta^{(N_h-1)}\right) \circ \ldots \circ \mathbf{h}^{(0)}\left(\theta^{(0)}\right)\right)(\mathbf{x})$



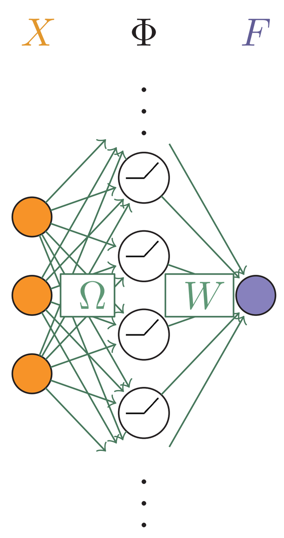
- Inference requires calculating the marginal likelihood:

$$p(X|\theta) = \int p\left(X|F^{(N_L)}, \theta^{(N_L)}\right) \times p\left(F^{(N_L)}|F^{(N_L-1)}, \theta^{(N_L-1)}\right) \times \ldots \times p\left((F^{(1)}|F^{(N_0)}, \theta^{(0)}\right) dF^{(N_L)} \ldots dF^{(1)}$$

DGP-AEs with Random Features

- GPs are single layered Neural Nets with an infinite number of hidden units
- Weight-space view of a GP:

$$F = \Phi W$$



- The priors over the weights are:

$$p(W_{\cdot i}) = \mathcal{N}(\mathbf{0}, I)$$

- The **RBF kernel** can be approximated using **trigonometric functions**

$$\Phi_{\mathrm{RBF}} = \sqrt{\frac{\sigma^2}{N_{\mathrm{RF}}}}\left[\cos(F\Omega), \sin(F\Omega)\right] \quad \text{with} \quad p(\Omega_{\cdot j}|\theta) = \mathcal{N}\left(\mathbf{0}, \Lambda^{-1}\right)$$

allowing for scaling factors σ^2 and $\Lambda = \mathrm{diag}(l_1^2, \ldots, l_d^2)$ for the kernel and the features (ARD);

- The first order **Arc-Cosine kernel** can be approximated using **Rectified Linear Units (ReLU)**

$$\Phi_{\mathrm{ARC}} = \sqrt{\frac{2\sigma^2}{N_{\mathrm{RF}}}} \max(\mathbf{0}, F\Omega) \quad \text{with} \quad p(\Omega_{\cdot j}|\theta) = \mathcal{N}\left(\mathbf{0}, \Lambda^{-1}\right)$$

- DGP-AEs with RFs become DNNs with low-rank weight matrices

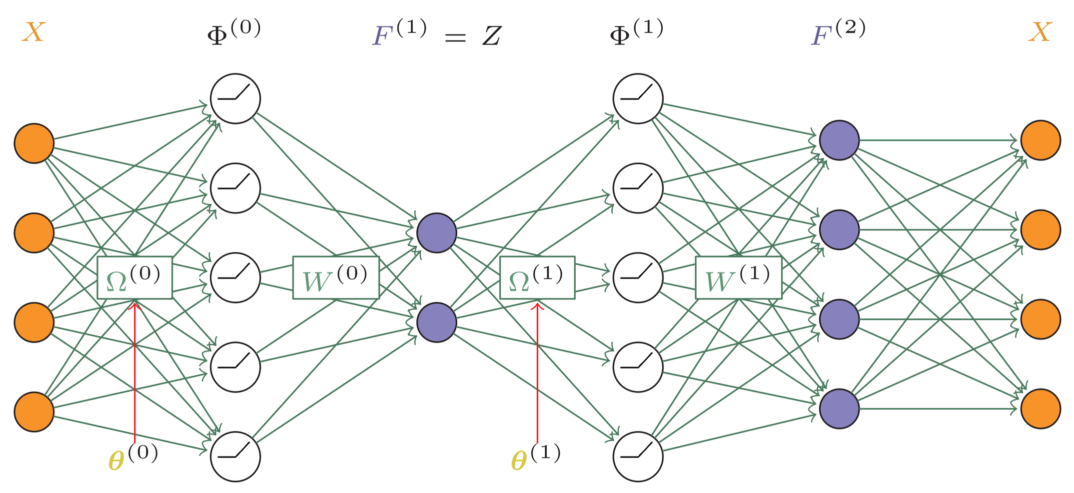


Fig. 1: Diagram of the proposed DGP autoencoder with random features (2 layers).

Stochastic Variational Inference

- Define $\Psi = \left(\Omega^{(0)}, \ldots, \Omega^{(L)}, W^{(0)}, \ldots, W^{(L)}\right)$
- Lower bound on the marginal likelihood:

$$\log\left[p(X|\theta)\right] \geq \mathbb{E}_{q(\Psi)}\left(\log\left[p\left(X|\Psi, \theta\right)\right]\right) - \mathrm{D_{KL}}\left[q(\Psi) \| p\left(\Psi\right)\right]$$

where $q(\Psi)$ approximates $p(\Psi|X, \theta)$

- Factorized approximate posterior: $q(\Psi) = \prod_{ijl} q\left(\Omega_{ij}^{(l)}\right) \prod_{ijl} q\left(W_{ij}^{(l)}\right)$

with $q\left(W_{ij}^{(l)}\right) = \mathcal{N}\left(\mu_{ij}^{(l)}, (\sigma^2)_{ij}^{(l)}\right)$ and $q\left(\Omega_{ij}^{(l)}\right) = \mathcal{N}\left(m_{ij}^{(l)}, (s^2)_{ij}^{(l)}\right)$

- Assuming factorized likelihood, we can use **mini-batch** stochastic gradient optimization:

$$\mathbb{E}_{q(\Psi)}\left(\log\left[p(X|\Psi, \theta)\right]\right) \approx \frac{n}{m} \sum_{\mathrm{k} \in \mathcal{I}_m} \mathbb{E}_{q(\Psi)}\left(\log[p(\mathbf{x}_\mathrm{k}|\Psi, \theta)]\right)$$

- The expectation can be estimated using **Monte Carlo sampling**, with $\tilde{\Psi}_r \sim q(\Psi)$:

$$\mathbb{E}_{q(\Psi)}\left(\log\left[p(x_\mathrm{k}|\Psi, \theta)\right]\right) \approx \frac{1}{N_{\mathrm{MC}}} \sum_{r=1}^{N_{\mathrm{MC}}} \log[p(x_\mathrm{k}|\tilde{\Psi}_r, \theta)]$$

- Predictive distribution

$$p(\mathbf{x}_*|X, \theta) = \int p(\mathbf{x}_*|\Psi, \theta) p(\Psi|X, \theta) d\Psi \approx \int p(\mathbf{x}_*|\Psi, \theta) q(\Psi) d\Psi \approx \frac{1}{N_{\mathrm{MC}}} \sum_{r=1}^{N_{\mathrm{MC}}} p(\mathbf{x}_*|\tilde{\Psi}_r, \theta)$$

Experimental setup and results

- Novelty detection performance: DGP-AE shows the best novelty detection abilities.
- Combined likelihoods increase the performance on mixed-type features.

	DGP-AE G-1	DGP-AE G-2	DGP-AE GS-1	DGP-AE GS-2	VAE-DGP-2	AE-1	AE-5	VAE-1	VAE-2	NADE-2	RKDE	IFOREST
MAMMOGRAPHY	**0.222**	0.183	**0.222**	0.183	**0.221**	0.118	0.075	0.119	0.148	0.193	**0.231**	**0.244**
MAGIC-GAMMA-SUB	0.260	0.340	0.260	0.340	0.235	0.253	0.125	0.230	0.305	**0.398**	**0.402**	0.290
WINE-QUALITY	**0.224**	**0.203**	**0.224**	**0.203**	0.075	0.106	0.042	0.064	0.124	0.102	0.051	0.059
MUSHROOM-SUB	0.811	0.677	**0.940**	0.892	0.636	0.725	0.331	0.758	0.479	0.596	0.839	0.546
CAR	0.050	0.061	0.043	0.067	0.045	0.044	0.032	**0.071**	0.050	0.030	0.034	0.041
GERMAN-SUB	0.066	0.077	**0.106**	0.098	**0.113**	0.065	**0.103**	**0.104**	0.062	**0.118**	**0.109**	0.079
PNR	**0.190**	0.172	**0.190**	0.172	**0.201**	0.059	0.107	0.100	0.106	0.006	0.146	0.124
TRANSACTIONS	0.756	0.752	**0.810**	**0.835**	0.509	0.563	0.510	0.532	0.760	0.373	0.585	0.564
SHARED-ACCESS	0.692	0.738	0.692	0.738	0.668	0.546	**0.766**	0.471	0.527	0.239	**0.783**	0.746
PAYMENT-SUB	**0.173**	**0.173**	0.168	0.168	0.137	0.157	0.129	**0.175**	0.143	0.101	**0.180**	0.142
AIRLINE	**0.081**	**0.079**	**0.081**	**0.079**	0.060	0.063	0.059	0.068	0.074	0.064	-	0.069
AVERAGE	0.344	0.338	0.366	0.370	0.284	0.264	0.222	0.262	0.270	0.216	0.336	0.284

Table 1: Mean area under the precision-recall curve (MAP) per dataset and algorithm (5 runs). AIRLINE was excluded from the average.

- Network convergence: DGP-AE and variational autoencoders achieve the best likelihood.

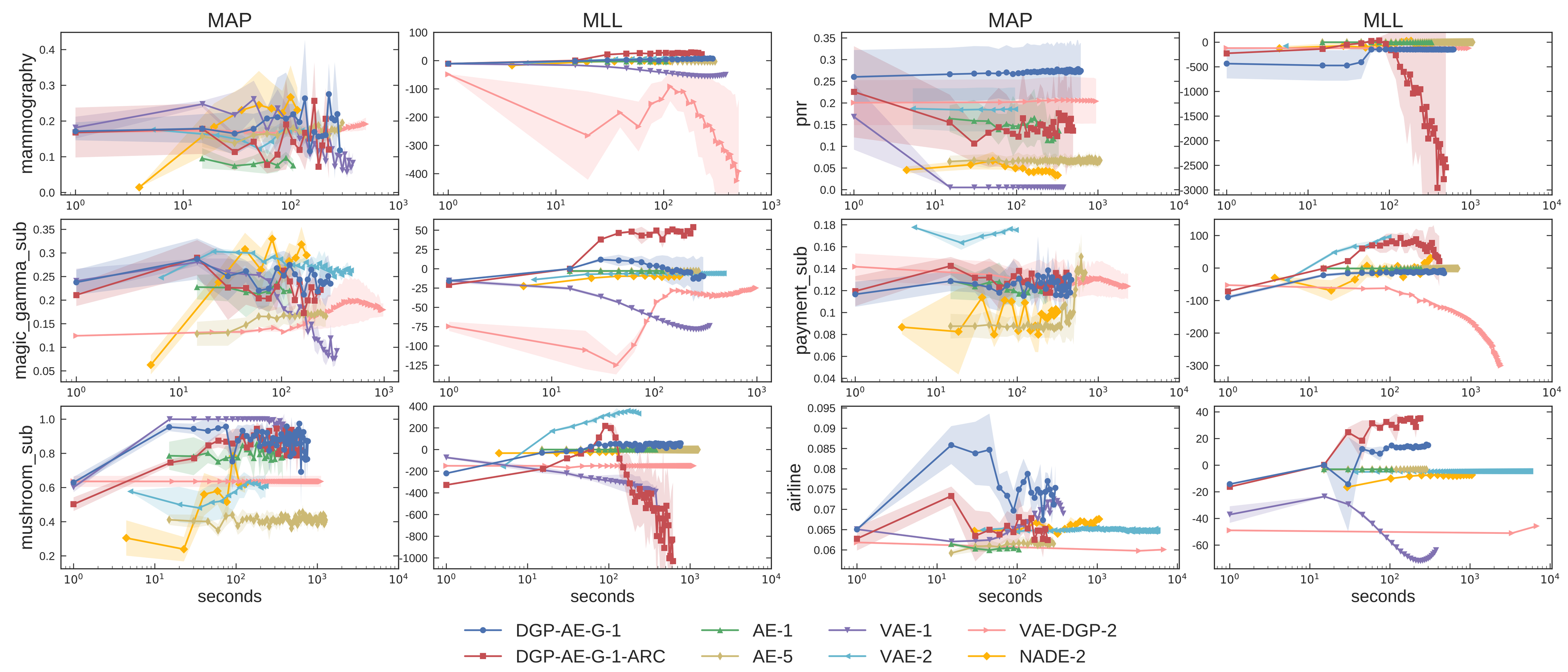


Fig. 2: Evolution of the MAP and MLL over time for the selected networks. Both metrics are computed on testing data. The higher values, the better the results.

- Depth: Moderately deep networks provide a high capacity and a fast convergence.

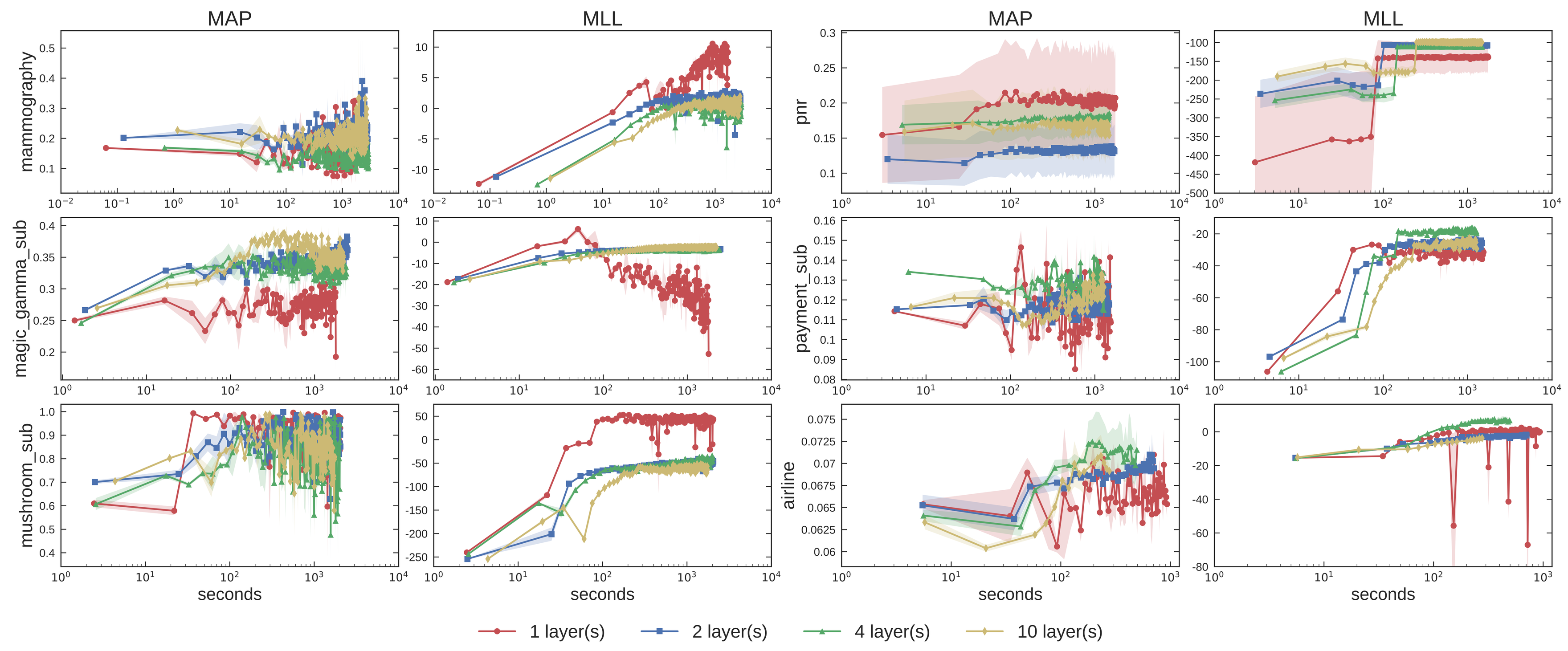
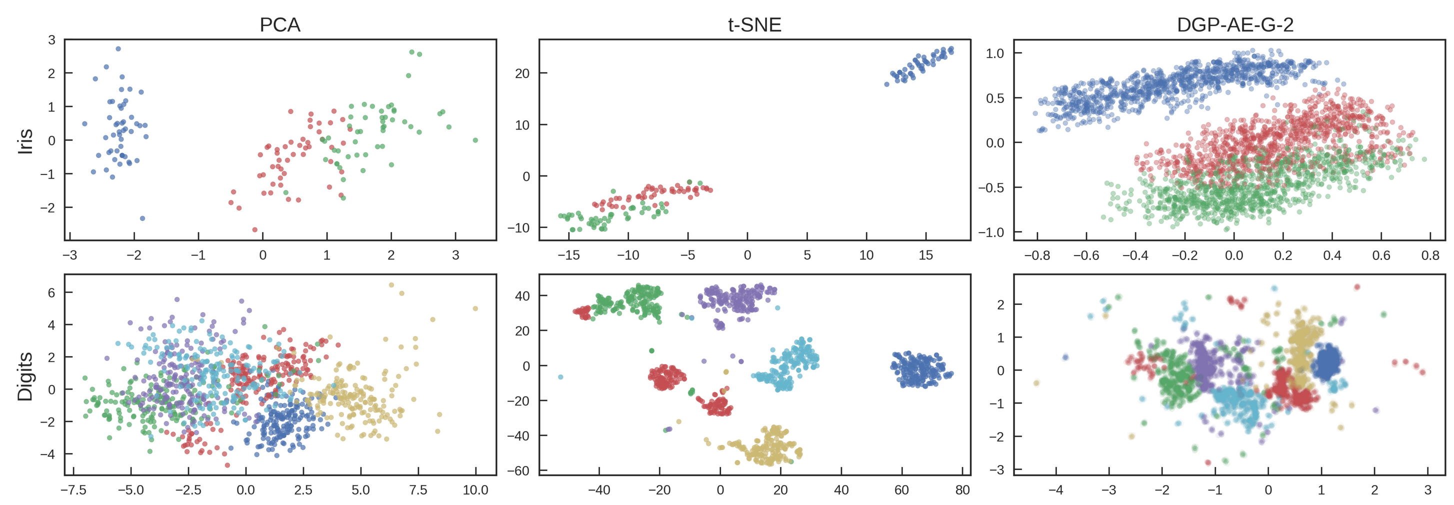


Fig. 3: Evolution of the MAP and MLL over time on testing data for DGP-AE with an increasing number of layers. For networks with more than 2 layers, we feed forward the input to the encoding layers, and feed forward the latent variables to the decoding layers. We use 3 gps per layer and a length-scale of 1.

- Dimensionality reduction: Meaningful low dimensional latent representations



Conclusions

- Contributions:
  - ✓ Novel deep probabilistic model for novelty detection
  - ✓ Competitive with state-of-the-art and DNN-based novelty detection methods
  - ✓ Good dimensionality reduction abilities
  - ✓ Can model mixed-types features
  - ✓ Tractable, scalable and suitable for distributed and GPU computing
- Future work:
  - Model discrete event sequences with structured DGP-AE
  - Generative DGP-AE

References

- R. Domingues, P. Michiardi, J. Zouaoui, and M. Filippone. Deep Gaussian Process autoencoders for novelty detection. *Machine Learning*, Jun 2018.
- Cutajar, K., Bonilla, E. V., Michiardi, P., & Filippone, M. (2017). Random feature expansions for Deep Gaussian Processes. In *Proceedings of the 34th international conference on machine learning, volume 70 of proceedings of machine learning research* (ICML 2017).
- Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74, 406-421.
- Dai, Z., Damianou, A., Gonzàlez, J., & Lawrence, N. (2016). Variationally auto-encoded Deep Gaussian Processes. In *Proceedings of the fourth international conference on learning representations (ICLR 2016)*.