Deep Gaussian Process Autoencoders for Novelty Detection

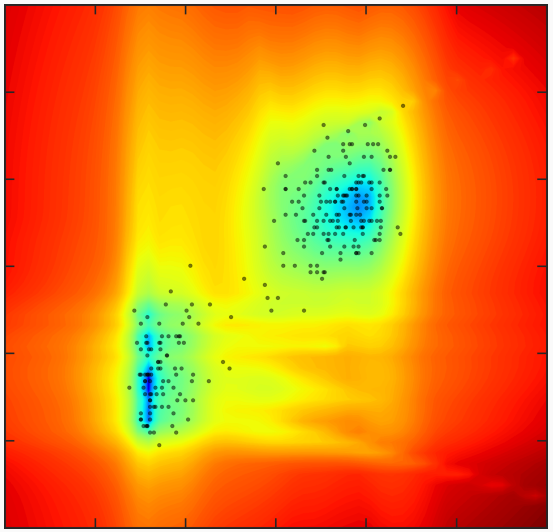
R. Domingues [1], **P. Michiardi** [1], **J. Zouaoui** [2], **M. Filippone** [1]

[1] EURECOM, Sophia Antipolis, France
[2] Amadeus, Sophia Antipolis, France

ECML-PKDD 2018 - Journal track
September 10th 2018

Novelty detection

- **Recognition of anomalies in test data which differ significantly from the training set**
- Estimate the distribution of nominal samples
- Similar to a one-class classification problem
- Usually addressed by unsupervised learning
- Training set may be contaminated by outliers



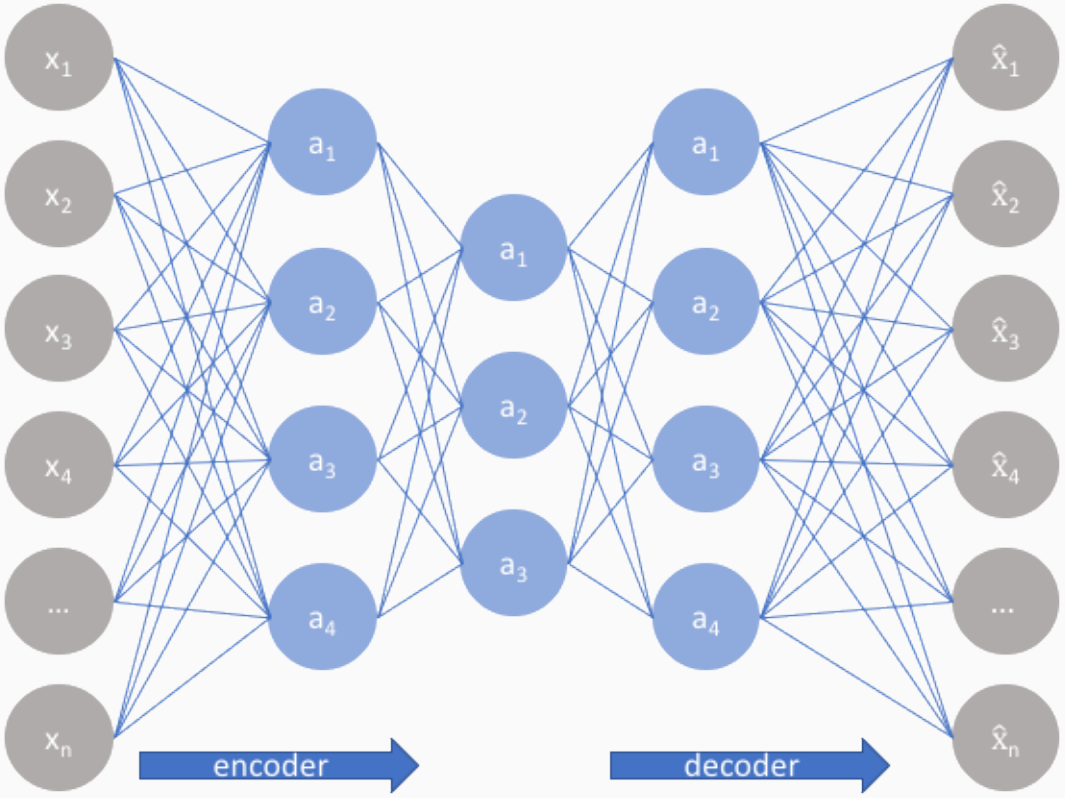
Deep Gaussian Process autoencoder

- Unsupervised and probabilistic
- Suitable for any type of data
- Training only requires tensor products
- Inference through stochastic variational inference
- Mini-batch learning

DGP-AE Architecture

Autoencoders

- Learn a compressed representation of the training data by minimizing the error between the input data and the reconstructed output

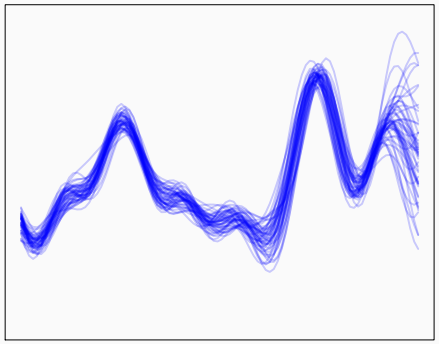


Deep Gaussian Process autoencoders

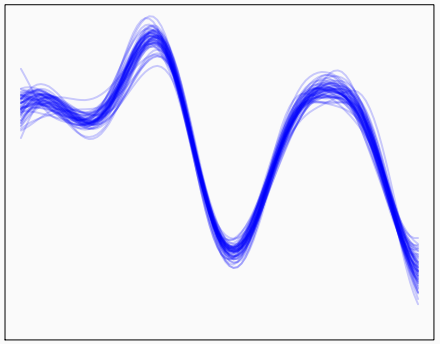
- Deep probabilistic models
- Composition of functions

$$\mathbf{f}(\mathbf{x}) = \left(\mathbf{h}^{(N_\mathrm{h}-1)}\left(\theta^{(N_\mathrm{h}-1)}\right) \circ \ldots \circ \mathbf{h}^{(0)}\left(\theta^{(0)}\right)\right)(\mathbf{x})$$

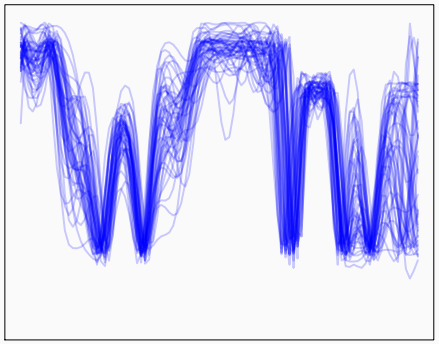
$\mathbf{h}^{(0)}(\mathbf{x})$



$\mathbf{h}^{(1)}(\mathbf{x})$



$\mathbf{h}^{(1)}\left(\mathbf{h}^{(0)}(\mathbf{x})\right)$



Marginal likelihood

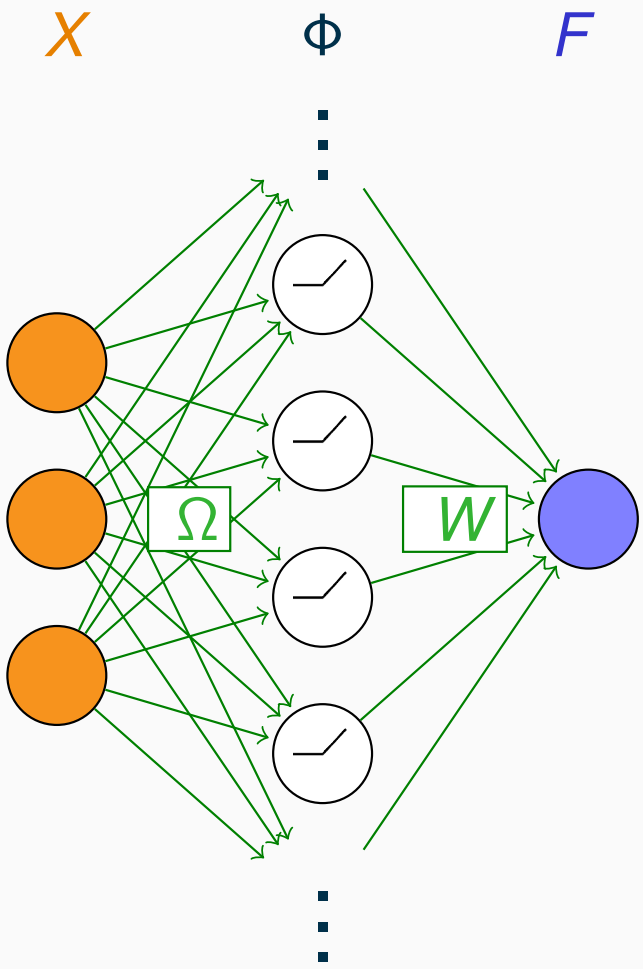
- Inference requires calculating the marginal likelihood:

$$
\begin{aligned}
p(X|\theta) \quad = \quad & \int p\left(X|F^{(N_\mathrm{L})}, \theta^{(N_\mathrm{L})}\right) \times \\
& p\left(F^{(N_\mathrm{L})}|F^{(N_\mathrm{L}-1)}, \theta^{(N_\mathrm{L}-1)}\right) \times \ldots \times \\
& p\left(F^{(1)}|F^{(N_0)}, \theta^{(0)}\right) dF^{(N_\mathrm{L})} \ldots dF^{(1)}
\end{aligned}
$$

Model inference

DGPs 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

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

Random Feature Expansion of Kernels

- Low-rank approximation of GP covariance functions
- The **RBF kernel** can be approximated using **trigonometric functions**

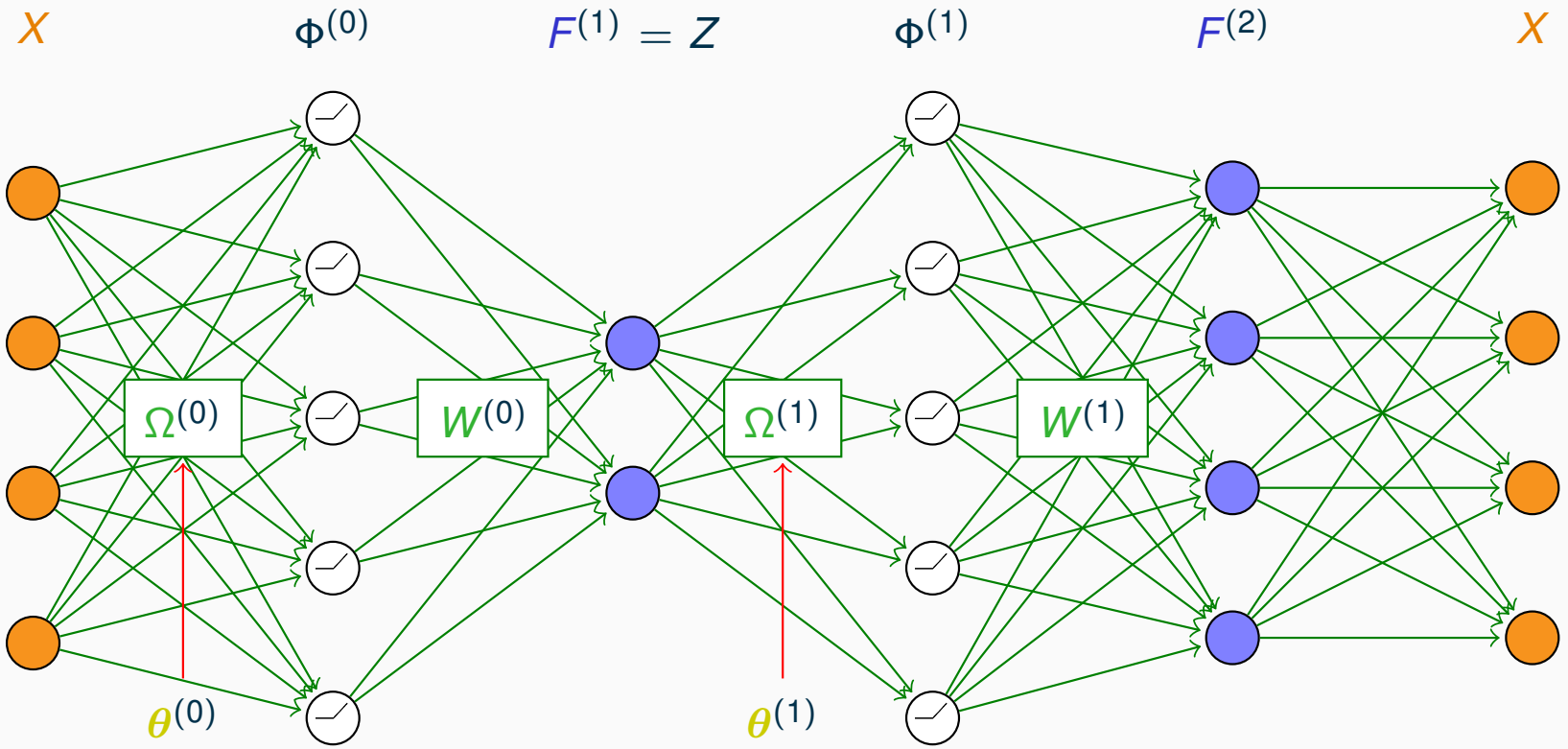
$$\Phi_{\text{RBF}} = \sqrt{\frac{\sigma^2}{N_{\text{RF}}}} \left[\cos\left(F\Omega\right), \sin\left(F\Omega\right)\right] \quad \text{with} \quad p\left(\Omega_{\cdot j} \middle| \theta\right) = \mathcal{N}\left(\mathbf{0}, \Lambda^{-1}\right)$$

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

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

- Approximated multivariate GPs are **Bayesian linear models**

DGP-AEs with RFs (2 layers)



DGP-AE with RFs - Stochastic Variational Inference

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

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

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

- D_{KL} computable analytically if q and p are Gaussian
- We assume an approximate factorized Gaussian distribution $q(\Psi)$

DGPs with RFs - Stochastic variational inference

- Stochastic **unbiased** estimate of the expectation term
  - **Mini-batch**

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

DGPs with RFs - Stochastic variational inference

- Stochastic **unbiased** estimate of the expectation term
  - **Mini-batch**

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

  - **Monte Carlo sampling**

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

    with $\tilde{\Psi}_r \sim q(\Psi)$

- The derivative of the estimate yields a **stochastic gradient**

DGPs with RFs - Stochastic variational inference

- Reparameterization trick

$$\left(\tilde{W}_r^{(l)}\right)_{ij} = s_{ij}^{(l)} \epsilon_{rij}^{(l)} + m_{ij}^{(l)}$$

with $\epsilon_{rij}^{(l)} \sim \mathcal{N}(0, 1)$

Predictions

- Predictive distribution

$$p(\mathbf{x}_*|X,\theta) = \int p(\mathbf{x}_*|\Psi,\theta)p(\Psi|X,\theta)d\Psi$$

- Approximation

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

Likelihood functions

- Model inference for mixed-type features
  - Normal: $p(x_{[G]}|\mathbf{f}^{(N_\mathrm{L})}) = \mathcal{N}(x_{[G]}|f_{[G]}^{(N_\mathrm{L})}, \sigma_{[G]}^2)$
  - Softmax: $p((\mathbf{x}_{[C]})_j|\mathbf{f}^{(N_\mathrm{L})}) = \frac{\exp[(f_{[C]}^{(N_\mathrm{L})})_j]}{\sum_i \exp[(f_{[C]}^{(N_\mathrm{L})})_i]}$
  - Combined likelihood: $p(\mathbf{x}|\mathbf{f}^{(N_\mathrm{L})}) = \prod_k p(\mathbf{x}_{[k]}|\mathbf{f}^{(N_\mathrm{L})})$

Evaluation

Algorithms

- **Isolation Forest:** IFOREST (Liu et al. 2008)
- **Robust Kernel Density Estimation:** RKDE (Kim and Scott 2012)
- **Feedforward Autoencoders:** AE-1, AE-5
- **Variational Autoencoders:** VAE-1, VAE-2 (Kingma and Welling 2014)
- **Variational Auto-Encoded DGP:** VAE-DGP-2 (Dai et al. 2016)
- **Neural Autoregressive Distribution Estimator:** NADE-2 (Uria et al. 2016)

Algorithms

- **Isolation Forest:** IFOREST (Liu et al. 2008)
- **Robust Kernel Density Estimation:** RKDE (Kim and Scott 2012)
- **Feedforward Autoencoders:** AE-1, AE-5
- **Variational Autoencoders:** VAE-1, VAE-2 (Kingma and Welling 2014)
- **Variational Auto-Encoded DGP:** VAE-DGP-2 (Dai et al. 2016)
- **Neural Autoregressive Distribution Estimator:** NADE-2 (Uria et al. 2016)

Algorithms

- **Isolation Forest:** IFOREST (Liu et al. 2008)
- **Robust Kernel Density Estimation:** RKDE (Kim and Scott 2012)
- **Feedforward Autoencoders:** AE-1, AE-5
- **Variational Autoencoders:** VAE-1, VAE-2 (Kingma and Welling 2014)
- **Variational Auto-Encoded DGP:** VAE-DGP-2 (Dai et al. 2016)
- **Neural Autoregressive Distribution Estimator:** NADE-2 (Uria et al. 2016)

Algorithms

- **Isolation Forest:** IFOREST (Liu et al. 2008)
- **Robust Kernel Density Estimation:** RKDE (Kim and Scott 2012)
- **Feedforward Autoencoders:** AE-1, AE-5
- **Variational Autoencoders:** VAE-1, VAE-2 (Kingma and Welling 2014)
- **Variational Auto-Encoded DGP:** VAE-DGP-2 (Dai et al. 2016)
- **Neural Autoregressive Distribution Estimator:** NADE-2 (Uria et al. 2016)

Algorithms

- **Isolation Forest:** IFOREST (Liu et al. 2008)
- **Robust Kernel Density Estimation:** RKDE (Kim and Scott 2012)
- **Feedforward Autoencoders:** AE-1, AE-5
- **Variational Autoencoders:** VAE-1, VAE-2 (Kingma and Welling 2014)
- **Variational Auto-Encoded DGP:** VAE-DGP-2 (Dai et al. 2016)
- **Neural Autoregressive Distribution Estimator:** NADE-2 (Uria et al. 2016)

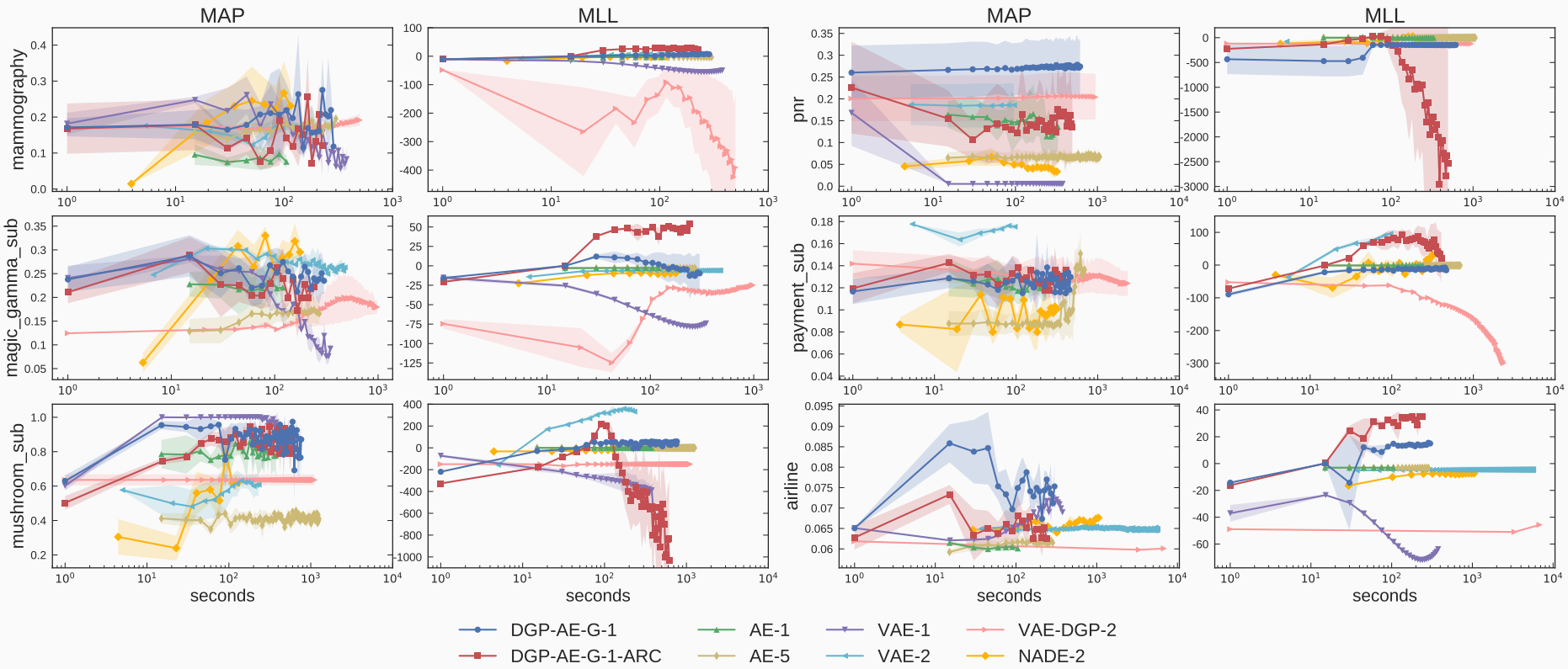
Method comparison

- 11 datasets, mean area under the precision-recall curve (MAP)
- Some datasets contain over 3 millions samples and 100 features
- DGP-AE achieves the best results for novelty detection
- Softmax accurately models categorical variables

	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

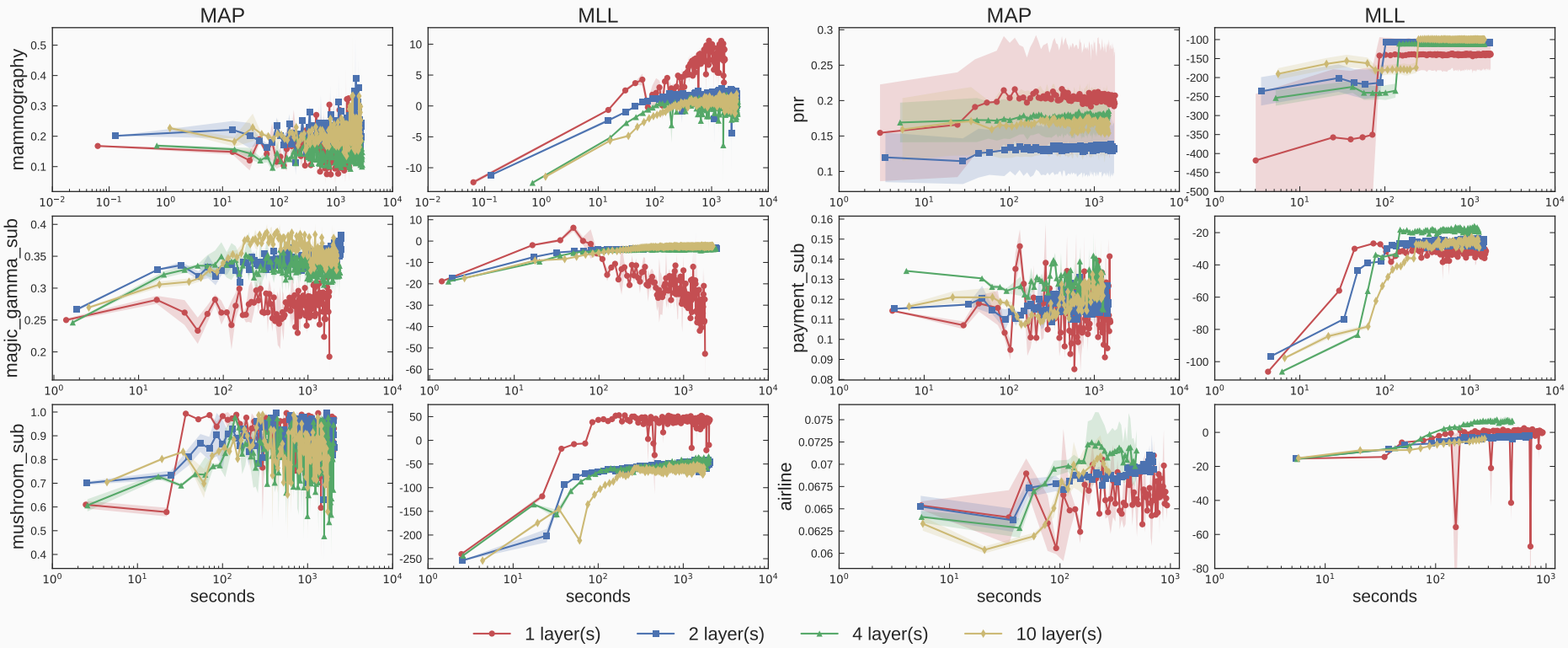
Convergence monitoring - Networks

- MAP and mean log-likelihood (MLL). The higher the better



- DGP-AE shows the best likelihood
- MAP quickly stabilizes while the likelihood is continuously refined

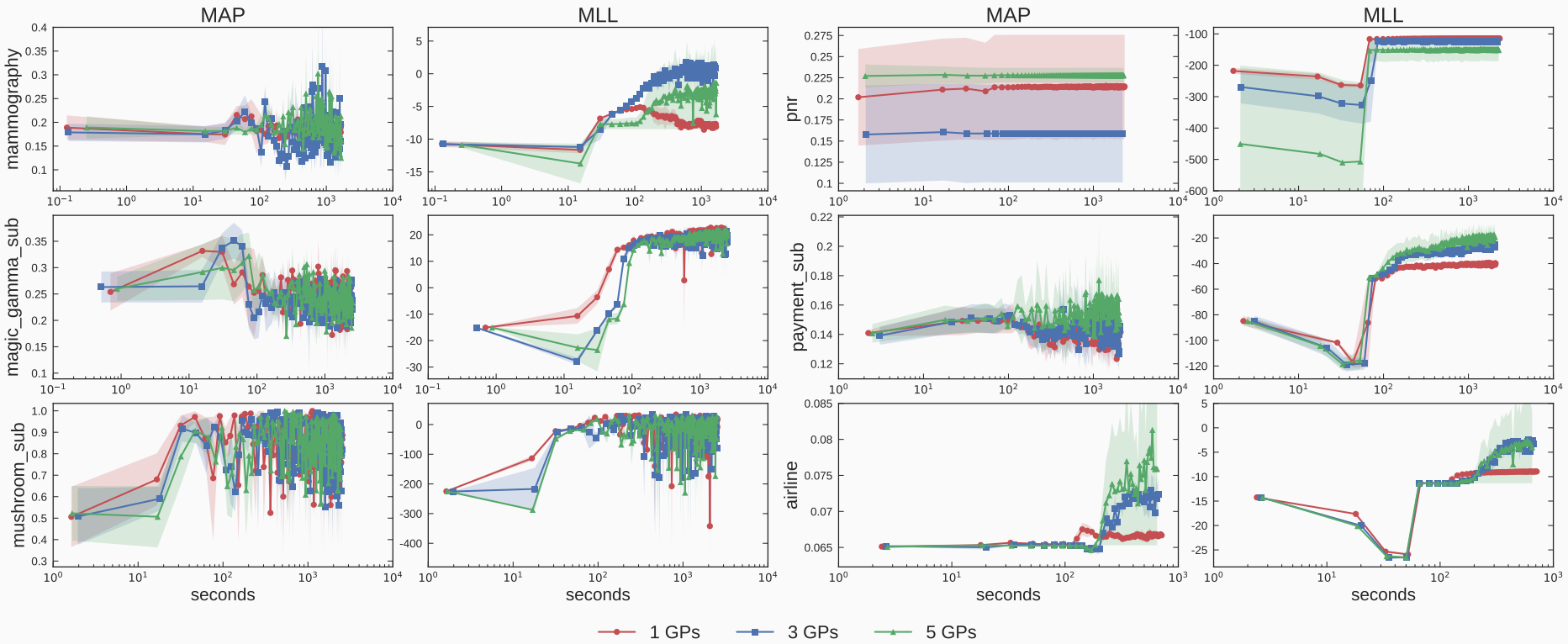
Convergence monitoring - Depth



- Correlation between a higher test likelihood and a higher MAP
- Moderately deep networks capture the complexity of data without an important convergence overhead

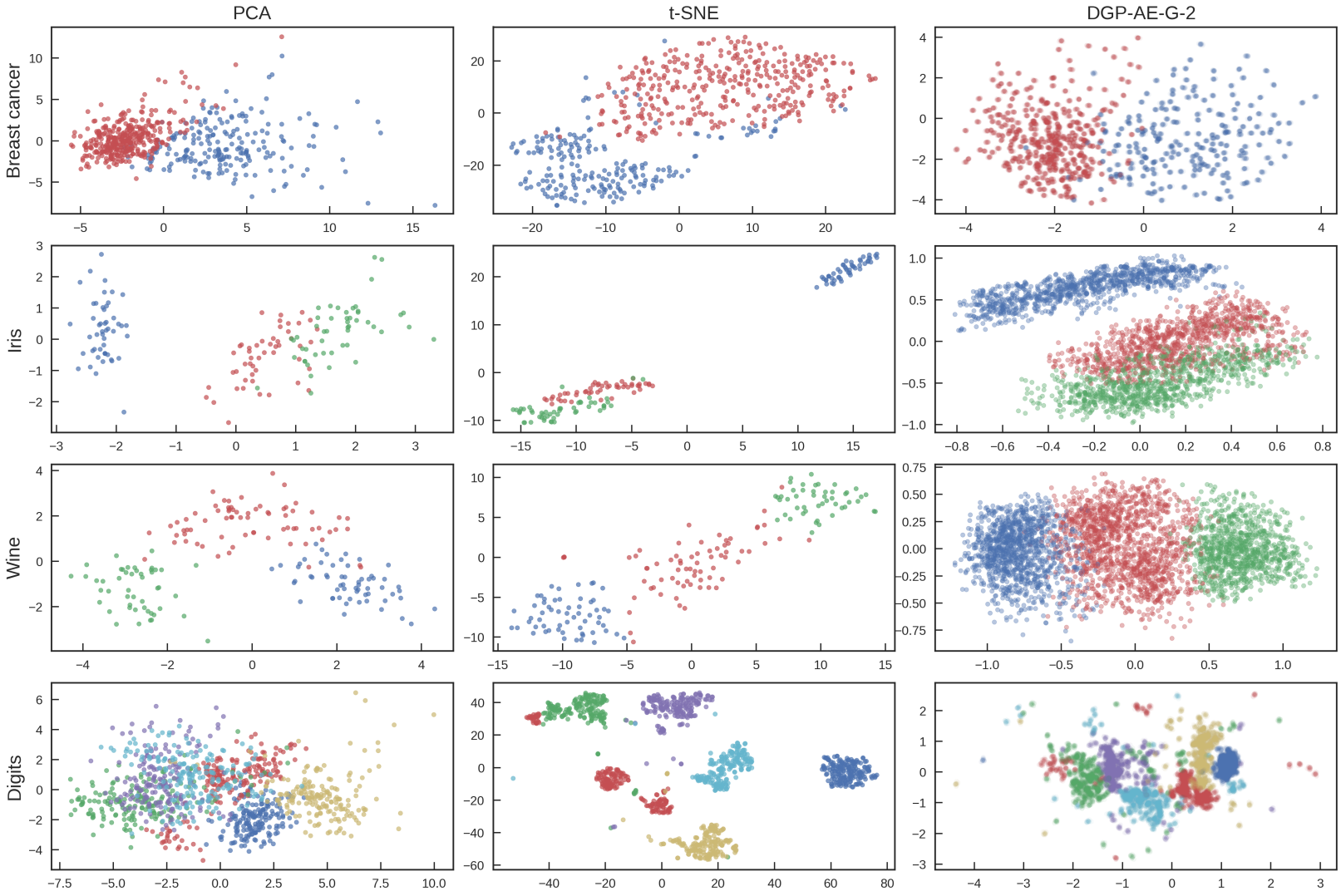
Convergence monitoring - GPs

- Dimensionality reduction capabilities of a DGP-AE-G-2



- Increasing the number of GPs results in a slower convergence
- 5 GPs achieve good novelty detection performance despite a significant dimensionality reduction

Latent representation



- Meaningful low-dimensional representations, comparable with state-of-the-art manifold learning methods

Conclusions

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
  - Tractable and scalable inference through random feature expansions and stochastic variational inference
  - Suitable to model mixed-types features

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
  - Tractable and scalable inference through random feature expansions and stochastic variational inference
  - Suitable to model mixed-types features

- **Future work**
  - Model discrete event sequences with structured DGP-AE
  - Generative DGP-AE

Thank you!