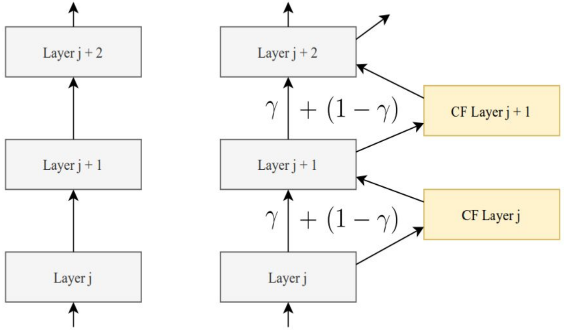
DONDERS INSTITUTE

Automatic variational inference with cascading flows

Luca Ambrogioni, Gianluigi Silvestri and Marcel van Gerven

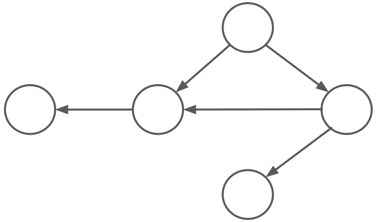
$$p(\boldsymbol{x}) = \prod_t \mathcal{N}(x_t; \mu(x_t), s^2(x_t))$$

$$q_{\boldsymbol{w}}(\boldsymbol{x}) = \prod_j^N \mathcal{T}_j^{\boldsymbol{w}} \left[\rho_j \left(\cdot \mid \theta_j(\boldsymbol{\pi}_j)\right)\right](x_j)$$



Radboud University Radboudumc

Differentiable probabilistic programming



Joint Distribution

Parents of j-th variable

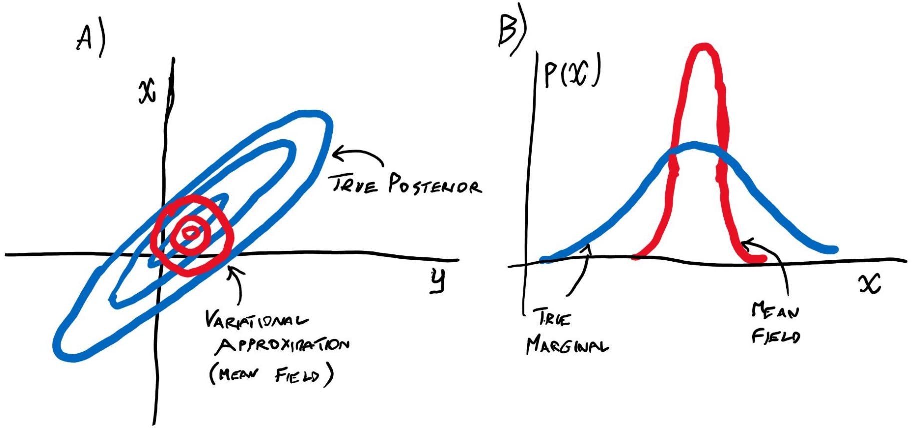
$$p(\boldsymbol{x}) = \prod_{j=1}^{N} \rho_j \left(x_j \mid \theta_j(\boldsymbol{\pi}_j) \right)$$

Re-parameterized distribution of j-th variable

Link function

VI performance depends on the variational parameterization

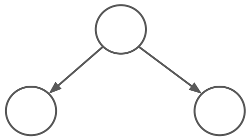
$$q(\boldsymbol{x}; \boldsymbol{\psi}) = \prod_k q_k(x_k; \psi_k)$$

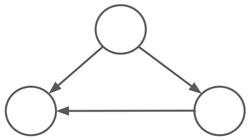


$$\nu = \sigma^2(1 - \rho^2)$$

Automatic construction of structured variational families

$$p(\boldsymbol{x}) \longmapsto q(\boldsymbol{x}; \psi)$$

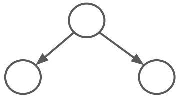




Automatic differentiation variational inference

Constrained variable

$$p\left(\boldsymbol{x}\right) = \prod_{j=1}^{N} \rho_j \left(x_j \mid \theta_j(\boldsymbol{\pi}_j)\right)$$

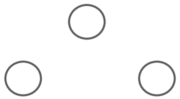


$f_j(x_j)$

$\boldsymbol{z}_j$

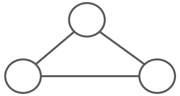
Unconstrained variable

Mean-field ADVI



$$q(\boldsymbol{x}) = \prod_j \left(\frac{\mathrm{d}f_j(x_j)}{\mathrm{d}x_j}\right)^{-1} \mathcal{N}(f_j^{-1}(x_j); \mu_k, s_k^2)$$

Multivariate normal ADVI



$$q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{f}^{-1}(\boldsymbol{x}); \boldsymbol{\mu}, HH^T) \prod_j \left(\frac{\mathrm{d}f_j(x_j)}{\mathrm{d}x_j}\right)^{-1}$$

Kucukelbir, Alp, et al. "Automatic differentiation variational inference." *The Journal of Machine Learning Research* 18.1 (2017): 430-474.

ASVI preserves the forward pass of the model

$$p(\boldsymbol{x}) = \prod_t \mathcal{N}(x_t; \mu(x_t), s^2(x_t))$$

$$q(\boldsymbol{x}) = \prod_t \mathcal{N}(x_t; \lambda_t^{\mu} \mu(x_t) + (1 - \lambda_t^{\mu})\alpha_t^{\mu}, \lambda_t^{s} s^2(x_t) + (1 - \lambda_t^{s})\alpha_t^{s})$$

Gate parameters

Perturbation parameters

Ambrogioni, Luca, et al. "Automatic structured variational inference." *AISTATS* (2021).

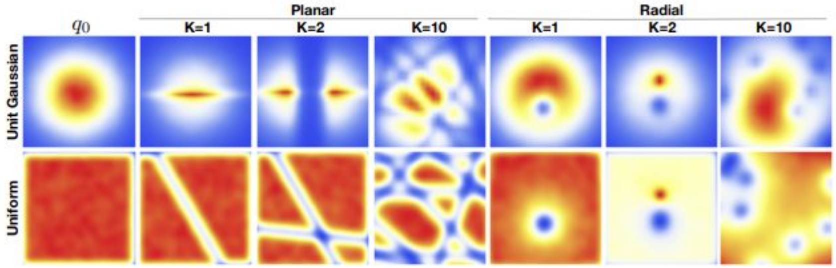
Variational inference with normalizing flows

Normal distribution

Non-linear transformation

$$p_X(x) = |\det J(\Psi^{-1}(x))| p_0(\Psi^{-1}(x))$$

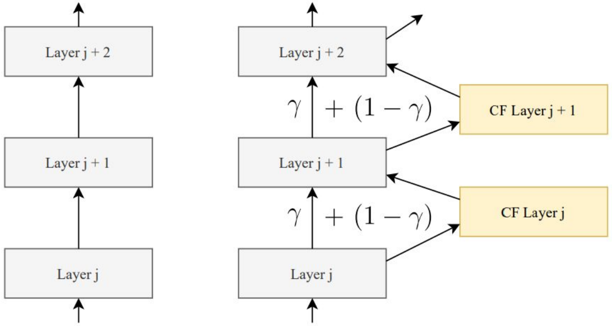
Volume distortion factor



Rezende, Danilo, and Shakir Mohamed. "Variational inference with normalizing flows." *International Conference on Machine Learning*. PMLR, 2015.

ASVI with cascading flows

$$p(\boldsymbol{x}) = \prod_t \mathcal{N}(x_t; \mu(x_t), s^2(x_t))$$



$$q_{\boldsymbol{w}}(\boldsymbol{x}) = \prod_j^N \mathcal{T}_j^{\boldsymbol{w}} \left[\rho_j\left(\cdot \mid \theta_j(\boldsymbol{\pi}_j)\right)\right](x_j)$$

Push-forward of non-linear transformation (normalizing flow)

Ambrogioni, Luca, Gianluigi Silvestri, and Marcel van Gerven. "Automatic variational inference with cascading flows." *arXiv preprint arXiv:2102.04801* (2021).

Highway flow architecture

1. Upper triangular highway layer:

$$l_U(z; U, \lambda) = \lambda z + (1 - \lambda)\,(Uz + b_U) \tag{9}$$

$$\log \det J_U = \sum_k \log\left(\lambda + (1 - \lambda) U_{kk}\right) \tag{10}$$

2. Lower triangular layer:

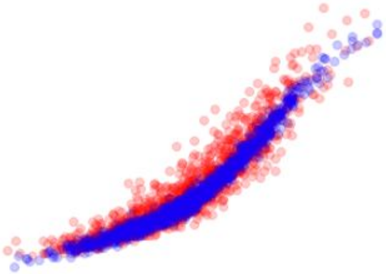
$$l_L(\boldsymbol{z}; L, \lambda) = \lambda z + (1 - \lambda)\,(Lz + b_L) \tag{11}$$

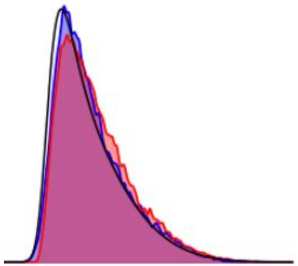
$$\log \det J_L = \sum_k \log\left(\lambda + (1 - \lambda) L_{kk}\right) \tag{12}$$

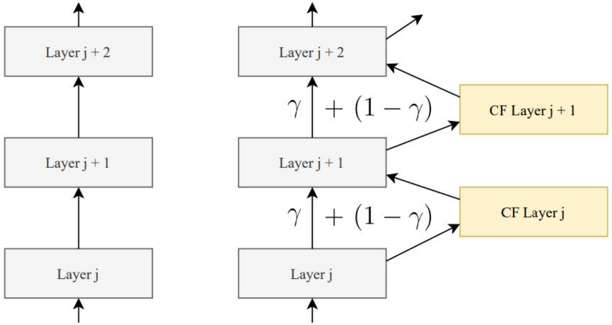
3. Highway activation functions:

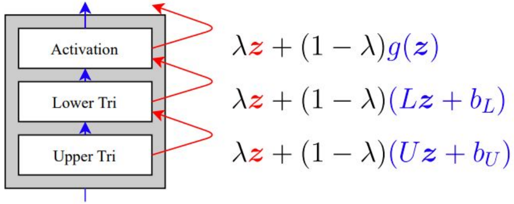
$$f(\boldsymbol{z}; \lambda) = \lambda z + (1 - \lambda) g(z) \tag{13}$$

$$\log \det \frac{\mathrm{d}f(x_k)}{\mathrm{d}x} = \sum_k \log\left(\lambda + (1 - \lambda)\frac{\mathrm{d}g(x_k)}{\mathrm{d}x}\right) \tag{14}$$









Ambrogioni, Luca, Gianluigi Silvestri, and Marcel van Gerven. "Automatic variational inference with cascading flows." *arXiv preprint arXiv:2102.04801* (2021).

Hierarchical variational inference and auxiliary variables

$$q(x_j, \epsilon_j \mid \boldsymbol{\pi}_j) = \hat{\mathcal{T}}_j^{\boldsymbol{w}} \left[\rho_j \left(\cdot \mid \theta_j(\boldsymbol{\pi}_j)\right) p_j(\cdot)\right] (x_j, \epsilon_j)$$

$$q(x_j \mid \boldsymbol{\pi}_j) = \int q(x_j, \boldsymbol{\epsilon}_j \mid \boldsymbol{\pi}_j)\, \mathrm{d}\boldsymbol{\epsilon}_j$$

Ranganath, Rajesh, Dustin Tran, and David Blei. "Hierarchical variational models." *International Conference on Machine Learning*. PMLR, 2016.

Caterini, Anthony, et al. "Variational Inference with Continuously-Indexed Normalizing Flows." *arXiv preprint arXiv:2007.05426* (2020).

Ambrogioni, Luca, Gianluigi Silvestri, and Marcel van Gerven. "Automatic variational inference with cascading flows." *arXiv preprint arXiv:2102.04801* (2021).

Hierarchical variational inference and auxiliary variables

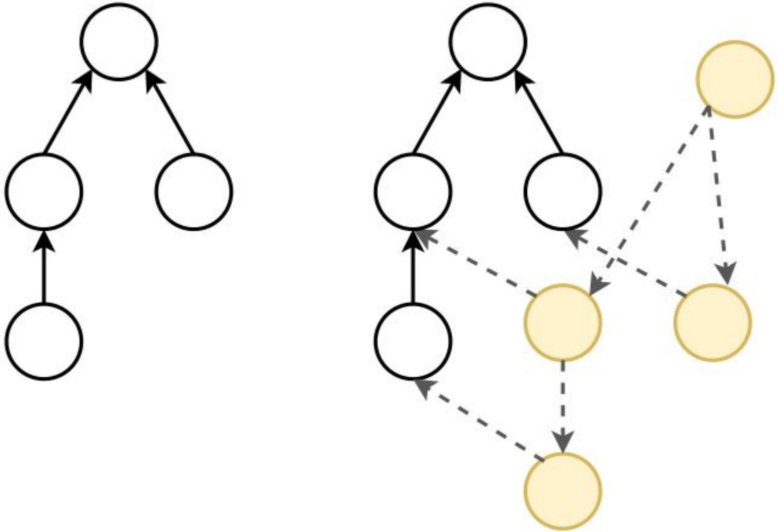
$$q\left(x_j \mid \boldsymbol{\pi}_j\right)=\int q\left(x_j, \boldsymbol{\epsilon}_j \mid \boldsymbol{\pi}_j\right) \mathrm{d} \boldsymbol{\epsilon}_j$$

$$\mathbb{E}_{\boldsymbol{x}}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{\int q(\boldsymbol{x}, \boldsymbol{\epsilon}) \mathrm{d} \boldsymbol{\epsilon}}\right] \geq \underbrace{\mathbb{E}_{\boldsymbol{x}, \boldsymbol{\epsilon}}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y}) r(\boldsymbol{\epsilon})}{q(\boldsymbol{x}, \boldsymbol{\epsilon})}\right]}_{\text{Augmented ELBO}}$$

Ranganath, Rajesh, Dustin Tran, and David Blei. "Hierarchical variational models."
International Conference on Machine Learning. PMLR, 2016.

Backward coupling and amortization

$$\epsilon_k \mid \boldsymbol{v}_k = \mathcal{B}^{(k)}[y_k] + \sum_{j=1}^{K} a_j \odot v_j + a_0 \odot \xi_k$$



Ambrogioni, Luca, Gianluigi Silvestri, and Marcel van Gerven. "Automatic variational inference with cascading flows." *arXiv preprint arXiv:2102.04801* (2021).

Experimental results

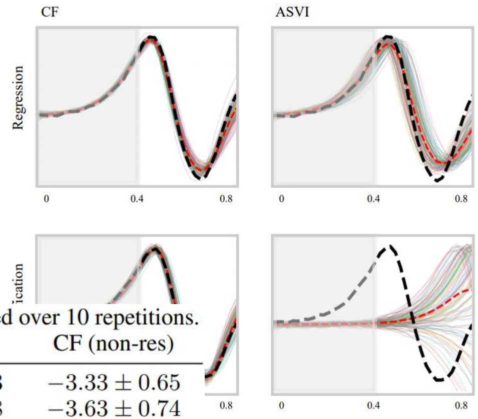


Table 1. Predictive and latent log-likelihood (forward KL) of variational timeseries models. Error are SEM estimated over 10 repetitions.

		CF	ASVI	MF	GF	MVN	CF (non-res)
BR-r	Pred	-2.27 ± 0.26	$\mathbf{-2.23 \pm 0.21}$	-3.79 ± 0.82	-2.81 ± 0.56	-2.88 ± 0.53	-3.33 ± 0.65
	Latent	-1.48 ± 0.19	$\mathbf{-1.45 \pm 0.14}$	-4.02 ± 0.63	-2.41 ± 0.52	-2.02 ± 0.48	-3.63 ± 0.74
BR-c	Pred	1.61 ± 0.18	1.45 ± 0.14	1.04 ± 0.03	$\mathbf{2.00 \pm 0.29}$	1.02 ± 0.03	1.31 ± 0.18
	Latent	$\mathbf{-1.53 \pm 0.21}$	-1.55 ± 0.19	-5.78 ± 0.89	-2.06 ± 0.53	-2.82 ± 0.77	-5.07 ± 0.85
LZ-r	Pred	$\mathbf{-2.89 \pm 0.17}$	-4.48 ± 0.60	-8.26 ± 0.28	-8.03 ± 0.37	-8.24 ± 0.29	-8.25 ± 0.27
	Latent	$\mathbf{-2.39 \pm 0.45}$	-4.38 ± 0.67	-10.28 ± 0.18	-9.44 ± 0.20	-9.45 ± 0.22	-10.00 ± 0.18
LZ-c	Pred	$\mathbf{5.10 \pm 0.52}$	0.92 ± 0.03	0.90 ± 0.003	0.86 ± 0.15	0.89 ± 0.001	0.88 ± 0.04
	Latent	$\mathbf{-4.19 \pm 0.66}$	-7.47 ± 0.30	-9.89 ± 0.19	-8.71 ± 0.32	-8.58 ± 0.34	-9.59 ± 0.29
PD-r	Pred	$\mathbf{-3.19 \pm 0.22}$	-3.25 ± 0.11	-4.42 ± 0.22	-3.84 ± 0.28	-4.30 ± 0.22	-4.29 ± 0.25
	Latent	$\mathbf{-2.32 \pm 0.19}$	-3.14 ± 0.12	-9.12 ± 0.29	-4.16 ± 0.33	-7.72 ± 0.30	-8.27 ± 0.36
PD-c	Pred	$\mathbf{1.97 \pm 0.07}$	1.65 ± 0.06	0.86 ± 0.003	01.07 ± 0.02	1.09 ± 0.02	0.96 ± 0.01
	Latent	$\mathbf{-2.77 \pm 0.18}$	-3.09 ± 0.15	-8.40 ± 0.43	-6.20 ± 0.40	-7.45 ± 0.42	-8.41 ± 0.43
RNN-r	Pred	-1.68 ± 0.05	-2.30 ± 0.18	-5.20 ± 0.94	$\mathbf{-1.60 \pm 0.09}$	-4.47 ± 0.92	-1.97 ± 0.21
	Latent	$\mathbf{-1.34 \pm 0.33}$	-1.95 ± 0.35	-10.30 ± 0.20	-6.39 ± 1.27	-6.61 ± 0.50	-10.47 ± 0.22
RNN-c	Pred	$\mathbf{5.77 \pm 1.40}$	1.05 ± 0.06	0.81 ± 0.03	2.81 ± 0.36	0.86 ± 0.02	1.39 ± 0.04
	Latent	-2.30 ± 0.61	$\mathbf{-2.05 \pm 0.32}$	-10.22 ± 0.29	-10.75 ± 0.15	-10.22 ± 0.29	-11.22 ± 0.04

Ambrogioni, Luca, Gianluigi Silvestri, and Marcel van Gerven. "Automatic variational inference with cascading flows." *arXiv preprint arXiv:2102.04801* (2021).

Radbound University / Donders Institute

Luca Ambrogioni

Gianluigi Silvestri

Marcel van Gerven