Copula-like Variational Inference

Marcel Hirt[1], Petros Dellaportas[1,2,3] and Alain Durmus[4]

University College of London, UK[1], Athens University of Economics and Business, Greece[2], The Alan Turing Institute, UK[3], CMLA, Ecole normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, France[4]

Introduction

- Variational inference aims at performing Bayesian inference by approximating an intractable posterior density π using some specified variational family $\mathbf{Q}$ of densities $(q_\xi)_{\xi\in\Xi}$, where the variational parameter is commonly chosen such that $\xi^\star \approx \arg\min_{\xi\in\Xi} \mathrm{KL}(q_\xi|\pi)$.
- Constructing an approximation family $\mathbf{Q}$ that is both flexible to closely approximate the density of interest and at the same time computationally efficient has been an ongoing challenge. One possibility is a Gaussian approximation with different types of covariance matrices such as diagonal matrices or low-rank perturbations thereof.
- Motivated by Sklar's theorem, variational families can be constructed using copula densities and one-dimensional marginal distributions as in [1] with a vine copula or as in [2] with a Gaussian copula.
- We propose simply to use a family of densities on the hypercube with non-uniform marginals that has linear complexity in the dimension of the state space for sampling and log-density evaluations.
- The flexibility of the variational distributions can be increased by applying sparse rotations as a novel normalizing flow [3] transformation.

Variational Inference and Copulas

- Consider Bayesian inference over latent variables $x \in \mathbb{R}^d$ having a prior density π_0 for a given likelihood function $L(y^{1:n}|x)$ with n observations $y^{1:n} = (y^1, \dots, y^n)$. The target density is the posterior $\pi(x) = p(x|y^{1:n})$ and minimizing $\xi \mapsto \mathrm{KL}(q_\xi|\pi)$ is equivalent to maximizing the so called ELBO (evidence lower bound)

$$\xi \mapsto \mathcal{L}(\xi) = \mathbb{E}_{q_\xi(x)}\left[\log \pi_0(x) + \log L(y^{1:n}|x) - \log q_\xi(x)\right]. \quad (1)$$

- To obtain more expressive variational distributions, samples from an initial density can be transformed through a sequence of invertible mappings $\{\mathscr{T}_t\}_{t=1}^T$, often termed normalizing flows [3]. Motivated by Sklar's theorem, one can choose as initial density any density c_θ with support on the hypercube $[0,1]^d$ that does not necessarily induce uniform marginals as any copula density would, and then apply the transformation $\mathscr{G}\colon [0,1]^d \to \mathbb{R}^d$, for any choice of cumulative distribution functions (cdfs) F_i, $i \in \{1, \dots, d\}$, with

$$\mathscr{G}\colon u \mapsto (F_1^{-1}(u_1), \dots F_d^{-1}(u_d)). \quad (2)$$

Basis Copula-like Density Function

The family of copula-like density that we consider is given by

$$c_\theta(v_1, \dots, v_d) = \frac{\Gamma(\alpha^*)}{\mathrm{B}(a,b)} \left[\prod_{\ell=1}^d \left\{\frac{v_\ell^{\alpha_\ell - 1}}{\Gamma(\alpha_\ell)}\right\}\right] (v^*)^{-\alpha^*} \quad (3)$$

$$\cdot \left(\max_{i\in\{1,\dots,d\}} v_i\right)^a \left[\left(1 - \max_{i\in\{1,\dots,d\}} v_i\right)^{b-1}\right],$$

with $v^* = \sum_{i=1}^d v_i$, $\alpha^* = \sum_{i=1}^d \alpha_i$ and $\theta = (a, b, (\alpha_i)_{i\in\{1,\dots,d\}}) \in (\mathbb{R}_+^* \times \mathbb{R}_+^* \times (\mathbb{R}_+^*)^d) = \Theta$.

Copula-like distribution without rotations

- The following probabilistic construction can be shown to allow for efficient sampling from the proposed density: Let $\theta \in \Theta$ and suppose that
  1. $(W_1, \dots, W_d) \sim \mathrm{Dirichlet}(\alpha_1, \dots, \alpha_d)$;
  2. $G \sim \mathrm{Beta}(a, b)$;
  3. $(V_1, \dots, V_d) = (GW_1/W^*, \dots, GW_d/W^*)$, where $W^* = \max_{i\in\{1,\dots,d\}} W_i$.

  Then the distribution of $(V_1, \dots, V_d)$ has a density given by (3).
- The random variable $Y = GW$ in this construction has a Beta-Liouville [4] law, which allows to account for negative dependence, inherited from the Dirichlet distribution through a Beta stick-breaking construction, as well as positive dependence via a common Beta-factor.
- In high dimensions, the correlations of the samples $V \sim c_\theta$ tend to be non-negative. However, by transforming V using the operator

$$\mathscr{H}\colon v \mapsto (1-\delta)\,\mathrm{Id} + \{\mathrm{diag}(2\delta) - \mathrm{Id}\}v, \quad (4)$$

where Id is the identity operator and $\delta \in (0,1)^d$, one obtains a random variable $U = \mathscr{H}V$ with support within the hypercube which can have a more flexible dependence structure for appropriate δ. Note that $(\mathscr{H}v)_i = \delta_i v_i + (1-\delta_i)(1-v_i)$, so we end up with a convex combination of v_i and its antithetic version $1 - v_i$. We take initially at random $\delta \in [0,1]^d$ for the transformation $\mathscr{H}$ such that $\mathbb{P}(\delta_i = \epsilon) = p$ and $\mathbb{P}(\delta_i = 1-\epsilon) = 1-p$ with $p, \epsilon \in (0,1)$ (in our experiments $\epsilon = 0.01$ and $p = 1/2$) and keep it fixed.
- We call the random variable $X' = \mathscr{G}(U)$ a sample from the copula-like distribution, where $\mathscr{G}$ is defined in (2) with F_i the cdf of a Gaussian.

Copula-like distribution with rotations

- For an invertible mapping $\mathscr{T}$, a sample from the final variational distribution can be obtained by setting $X = \mathscr{T}X'$.
- We propose a new volume-preserving transformation $\mathscr{T}$ that is given as a rotation matrix $\mathcal{R}_d$ that can be represented as a product of $d/2 \log d$ Givens rotations, following the FFT-style butterfly-architecture proposed in [5]. It allows for a minimal number of rotations so that all d coordinates can interact with one another at a complexity of $\mathcal{O}(d \log d)$. For the case $d = 4$ say, the rotation matrix is

$$\mathcal{R}_4 = \begin{bmatrix} c_1 & -s_1 & 0 & 0 \\ s_1 & c_1 & 0 & 0 \\ 0 & 0 & c_3 & -s_3 \\ 0 & 0 & s_3 & c_3 \end{bmatrix} \begin{bmatrix} c_2 & 0 & -s_2 & 0 \\ 0 & c_2 & 0 & -s_2 \\ s_2 & 0 & c_2 & 0 \\ 0 & s_2 & 0 & c_2 \end{bmatrix} = \begin{bmatrix} c_1c_2 & -s_1c_2 & -c_1s_2 & s_1s_2 \\ s_1c_2 & c_1c_2 & -s_1s_2 & -c_1s_s \\ c_3s_2 & -s_3s_2 & c_3c_2 & -s_3c_s \\ s_3s_2 & c_3s_2 & s_3c_2 & c_3c_2 \end{bmatrix},$$

for $d-1$ parameters $\nu_1, \nu_2, \nu_3 \in \mathbb{R}$ and $c_i = \cos(\nu_i)$, $s_i = \sin(\nu_i)$.

Optimizing the Variational Bound

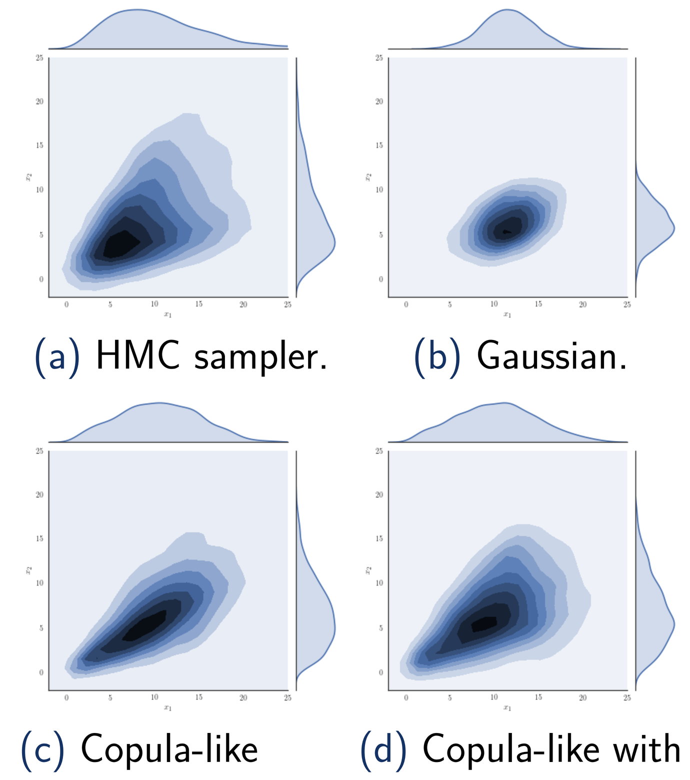
The density of the rotated variable $X = \mathcal{R}_d X'$ can be evaluated explicitly by using (3) together with the Jacobian-determinant of the used bijections. Stochastic gradients of (1) can then be obtained using an implicit reparametrization [6] for Dirichlet and Beta distributions.

Bayesian Logistic Regression

We compare variational families for a previously considered synthetic dataset using a logistic regression model in dimension $d = 2$ with a Gaussian prior.

Table: Comparison of the ELBO between different variational families for the logistic regression experiment.

Variational family	ELBO
Mean-field Gaussian	-3.42
Full-covariance Gaussian	-2.97
Copula-like without rotations	-2.30
Copula-like with rotations	-2.19



(a) HMC sampler. (b) Gaussian.

(c) Copula-like without rotation. (d) Copula-like with rotation.

Figure: Densities for logistic regression.

Bayesian Neural Network

We illustrate our approach for a fully-connected neural network with two hidden layers of size 200×200. We perform inference on MNIST and place Horseshoe priors on the weights, resulting in $d > 200{,}000$. As an ablation study, we also consider the case where the copula-like c_θ in (3) is replaced with an independent copula density, *i.e.* $c_\theta(v) = 1$ for $v \in [0,1]^d$. Additionally, we also analyse the case where the sparse rotation as the final transformation $\mathscr{T}\colon x' \mapsto \mathcal{R}_d x'$ is replaced by an affine autoregressive transformation [7], also known as an inverse autoregressive flow (IAF). We observe that the copula-like density has a better predictive performance compared to an independent copula. Further, applying the sparse rotation can be beneficial compared to the IAF for the copula-like density. Lastly, never using the antithetic component in (4) can limit the flexibility of the copula-like density.

Table: MNIST prediction errors on test set.

Variational approximation	Error Rate
Copula-like with rotations	1.70 %
Copula-like without rotations	1.78 %
Copula-like with IAF	2.04 %
Independent copula with IAF	2.88 %
Independent copula with rotations	2.90 %
Mean-field Gaussian	3.82 %
Copula-like without rotations and $\delta_i = 0.99$ for all i	5.70 %

References

[1] Tran *et al.*, *NIPS* 2015, 3564–3572. [2] Han *et al.*, *AISTATS* 2016, 829–838. [3] Rezende *et al.*, *ICML* 2014, 1278–1286. [4] Kai *et al.*, 2017. [5] Genz, 1998. [6] Figurnov *et al.*, *NeurIPS* 2018, 441-452. [7] Kingma *et al.*, *NIPS* 2016, 4743–4751.