Learning Doubly Intractable Latent Variable Models via Score Matching

Eszter Vértes, Maneesh Sahani

Gatsby Unit, University College London

Background

- Latent variable models are powerful tools for learning about the underlying structure of a dataset in an unsupervised setting
- Learning is intractable in most complex (e.g. non-Gaussian) models.

Double intractability:

1. The posterior distribution is intractable, i.e. we cannot compute the normalizer for the latent variables: $Z(\theta)_{z|x} = \int p(x,z)dz$
2. For some latent variable models the joint distribution is only available up to proportionality:

$$p(x,z) = \frac{1}{Z(\theta)}\tilde{p}(x,z), \text{ where } Z(\theta) = \int \tilde{p}(x,z)dxdz$$

→ Variational algorithms are infeasible since we do not have access to the normalized log-joint.

Score matching (SM)

- Score matching (Hyvarinen, 2005): method for estimating non-normalised statistical models without latent variables
- The explicit score matching objective function:

$$J(\theta) = E_x\left[\|\partial_x \log p^*(x) - \partial_x \log p_\theta(x)\|^2\right]$$

Where $p^*(x)$ is the true density and $p_\theta(x)$ is the model density.

- Hyvarinen showed that it is equivalent to minimising the following cost function:

$$\tilde{J}(\theta) = E_x\left[\partial_x^2 \log p_\theta(x) + \frac{1}{2}\left(\partial_x \log p_\theta(x)\right)^2\right]$$

Note that:

- It depends on the true density $p(x)$ only through its expectation, which can be evaluated by summing over data samples
- The *score function*, $\partial_x \log p_\theta(x)$ does not depend on the unknown normalizer

Score matching for latent variable models

- For energy based models of the form: $p(x,z) \propto \exp(-E_\theta(x,z))$
- The score function can be expressed as an expectation:

$$\partial_x \log p_\theta(x) = \int p(z|x)(-\partial_x E(x,z))dz$$

- The score matching objective can be rewritten (Swersky et al., 2011):

$$J(\theta) = \sum_x \sum_i -\frac{1}{2}\langle \partial_{x_i} E_\theta(x,z)\rangle_{z|x}^2 + \langle (\partial_{x_i} E(x,z))^2\rangle_{z|x} - \langle \partial_{x_i}^2 E_\theta(x,z)\rangle_{z|x}$$

Exponential family

- Jointly exponential family model:

$$p(x,z) = \exp\left(\theta^T S(x,z) - A(\theta)\right)$$

where θ: natural parameter vector, $S(x,z)$: sufficient statistic

- Useful property: $\nabla_\theta A(\theta) = \langle S(x,z)\rangle_{x,z}$

References

1. Hyvarinen, Aapo. "Estimation of non-normalized statistical models by score matching." Journal of Machine Learning Research 6.Apr (2005): 695-709.
2. Swersky, Kevin, et al. "On autoencoders and score matching for energy based models." Proceedings of the 28th international conference on machine learning (ICML-11). 2011.
3. Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. arxiv preprint arxiv: 1111.4246. 2011.

SM for doubly intractable models

- The exact SM objective for jointly exponential family models:

$$J(\theta) = \sum_x \sum_i -\frac{1}{2}\left\langle \theta^T \partial_{x_i} S(x,z)\right\rangle_{z|x}^2 + \left\langle \left(\theta^T \partial_{x_i} S(x,z)\right)^2\right\rangle_{z|x} + \langle \theta^T \partial_{x_i}^2 S(x,z)\rangle_{z|x}$$

- We can propagate derivatives wrt. θ into the expectations without knowing the normaliser of $p(z|x)$ or $p(x,z)$ by using the property of exp. family:

$$\nabla_\theta \log p(z|x) = S(x,z) - \langle S(x,z)\rangle_{z|x}$$

- The posterior $p(z|x)$ appears in the resulting gradient $\nabla_\theta J(\theta)$ only in terms of its expectations.
- We approximate these integrals using a Hamiltonian Monte Carlo sampler (Hoffman et al., 2011)

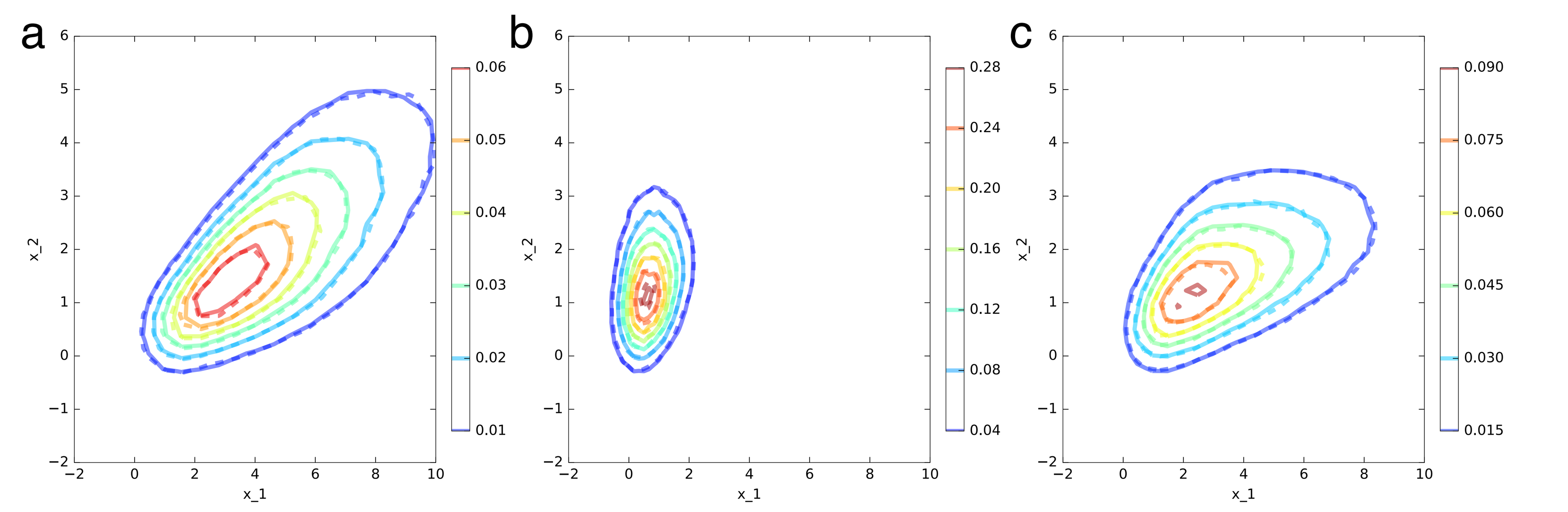
Experiments

- Rectified latent Gaussian model defined as:

$$p(z) \propto \mathcal{N}(z|\mathbf{0},\Sigma)\prod_l \Theta(z_l)$$
$$p(x|z) = \mathcal{N}(Wz, \sigma^2 I)$$

- Sufficient statistics: $S(x,z) = \mathbf{vec}\left[x^T x, xz^T, zz^T\right]$
- In general, the normalizer for the joint model cannot be computed analytically.
- $z \in \mathcal{R}_+^2, x \in \mathcal{R}^2$, we learn Σ, W, σ

Contours of learned and true densities

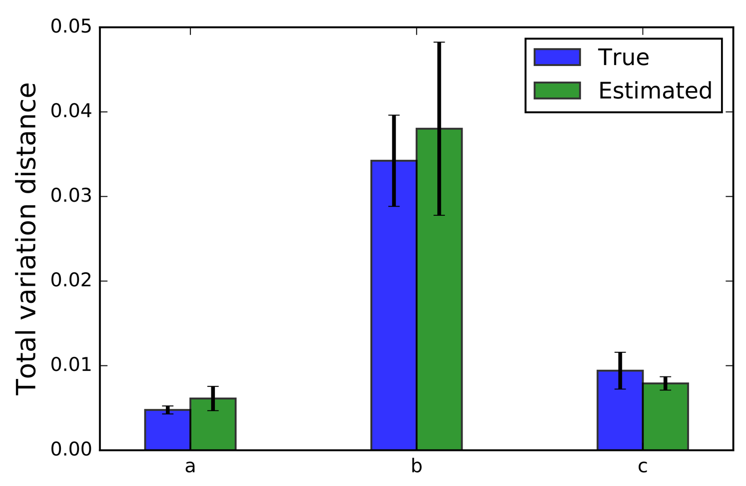


Total variation distance

- Empirical distance between two densities:

$$\delta(P,Q) = \sup_x |P(x) - Q(x)|$$

- Computed between pairs of data sets generated from the true and learned models (green) and between two data sets coming from the true model (blue)



Summary

- Score matching can be applied to doubly intractable jointly exponential family models
- SM allows for learning flexible latent variable models with arbitrary sufficient statistics
- No need for fixed form approximations of the posterior distribution
- In contrast to the Boltzmann machine learning rule or contrastive divergence, Monte Carlo simulation is only required for sampling from the posterior, not from the joint distribution

This work was funded by the Gatsby Charitable Foundation.

Contact: eszter@gatsby.ucl.ac.uk