Structured Variational Inference in Continuous Cox Process Models

Virginia Aglietti[1,2], Edwin V. Bonilla[3], Theodoros Damoulas[1,2], Sally Cripps[4]

[1] *University of Warwick* [2] *The Alan Turing Institute* [3] *CSIRO's Data61* [4] *The University of Sydney*

V.Aglietti@warwick.ac.uk, Edwin.Bonilla@data61.csiro.au, T.Damoulas@warwick.ac.uk, Sally.Cripps@sydney.edu.au

Contributions

We propose a scalable *structured* variational inference algorithm for *continuous* sigmoidal Cox processes (STVB). Contributions:

- **Scalable inference in continuous input spaces** via a process superposition.
- **Efficient structured posterior estimation** giving a posterior capturing the complex variable dependencies in the model
- **State-of-the-art performance** when compared to alternative inference schemes, link functions, augmentation schemes and representations of the input space.

The likelihood function

Discrete likelihood:

$$p(\mathbf{Y}|\mathbf{f}) = \prod_{n=1}^{N} \text{Poisson}\,(y_n; \lambda(\mathbf{x}))$$

Continuous likelihood:

$$\mathcal{L}(N, \{\mathbf{x}_1, ..., \mathbf{x}_n\}|\lambda(\mathbf{x})) = \exp\left(-\int_\tau \lambda(\mathbf{x})d\mathbf{x}\right) \prod_{n=1}^{N} \lambda(\mathbf{x}_n)$$

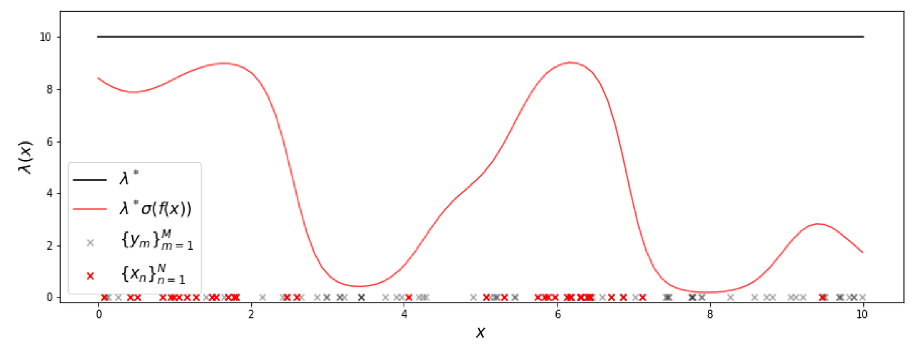


Figure 1: Superposition of two Poisson Point Processes with intensities $\lambda^\star\sigma(f(\mathbf{x}))$ and $\lambda^\star\sigma(-f(\mathbf{x}))$.

	STVB	LGCP	SGCP [1]	Gunter et al.(2014)	VBPP [2]	Lian et al. (2015)	MFVB [3]
Inference	SVI	MCMC	MCMC	MCMC	VI-MF	VI-MF	VI-MF
$\mathcal{O}$	K^3	N^3	$(N+M)^3$	$(N+M)^3$	NK^2	NK^2	NK^2
$\lambda(x)$	$\lambda^\star\sigma(f(x))$	$\exp(f(x))$	$\lambda^\star\sigma(f(x))$	$\lambda^\star\sigma(f(x))$	$(f(x))^2$	$(f(x))^2$	$\lambda^\star\sigma(f(x))$
$\mathcal{X}$	$\int$	$\sum$	$\int$	$\int$	$\int$	$\sum$	$\int$
Tractability	Superposition		Thinning	Adaptive Thinning	Functional form		Integral approximation

Table 1: Summary of related work. $\int$ is continuous, $\sum$ is discrete. M represents the number of thinned points derived from the thinning algorithm. K are the number of inducing inputs.

Augmentation via superposition

Full joint distribution $\mathcal{L}(\{\mathbf{x}_n\}_{n=1}^N, \{\mathbf{y}_m\}_{m=1}^M, M, \mathbf{f}, \lambda^\star|\mathcal{X}, \boldsymbol{\theta})$:

$$\frac{(\lambda^\star)^{N+M}\exp(-\lambda^\star\int_{\mathcal{X}} d\mathbf{x})}{N!M!} \prod_{n=1}^{N}\sigma(f(\mathbf{x}_n)) \prod_{m=1}^{M}\sigma(-f(\mathbf{y}_m))p(\mathbf{f})p(\lambda^\star)$$

Structured Variational Inference

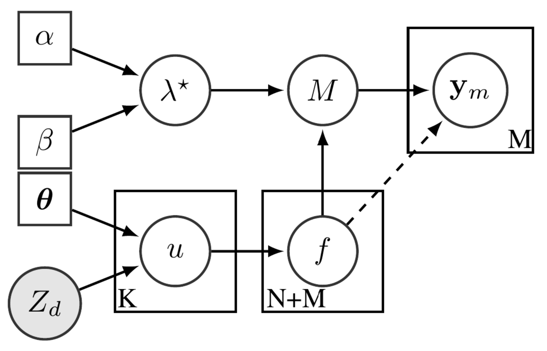


Figure 2: Posterior distribution accounting for all model dependencies. The dashed line represents the assumed factorization.

$$Q(\mathbf{f}, \mathbf{u}, M, \{\mathbf{y}_m\}_{m=1}^M, \lambda^\star) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})q(\lambda^\star) \times q(\{\mathbf{y}_m\}_{m=1}^M|M)q(M|\mathbf{f}, \lambda^\star)$$

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S}) \qquad q(\lambda^\star) = \text{Gamma}(\alpha, \beta)$$

$$q(\{\mathbf{y}_m\}_{m=1}^M|M) = \prod_{m=1}^{M}\sum_{s=1}^{S}\pi_s\mathcal{N}_T(\mu_s, \sigma_s^2; \mathcal{X})$$

$$q(M|\mathbf{f}, \lambda^\star) = \text{Poisson}(\eta) \qquad \eta = \lambda^\star\int_{\mathcal{X}}\sigma(-f(\mathbf{x}))d\mathbf{x}$$

The evidence lower bound

$$\mathcal{L}_{\text{elbo}} = T_0 + \underbrace{\mathbb{E}_Q[M\log(\lambda^\star)]}_{T_1} - \underbrace{\mathbb{E}_Q[\log(M!)]}_{T_2} + \sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{f})}[\log(\sigma(f(\mathbf{x}_n)))] + \underbrace{\mathbb{E}_Q\left[\sum_{m=1}^{M}\log(\sigma(-f(\mathbf{y}_m)))\right]}_{T_3} - \mathcal{L}_{\text{kl}}^{\mathbf{u}} - \mathcal{L}_{\text{kl}}^{\lambda^\star} - \underbrace{\mathcal{L}_{\text{ent}}^{M}}_{T_4} - \underbrace{\mathcal{L}_{\text{ent}}^{\{\mathbf{y}_m\}_{m=1}^M}}_{T_5}$$

where $T_0 = N(\psi(\alpha) - \log(\beta)) - V\frac{\alpha}{\beta} - \log(N!)$, $V = \int_{\mathcal{X}} d\mathbf{x}$, $\psi(\cdot)$ is the digamma function and $q(\mathbf{f}) = \mathcal{N}(\mathbf{Am}, \mathbf{K}_{xx} - \mathbf{AK}_{zx} + \mathbf{ASA}^T)$.

We derive expressions for $T_i, i = 1, ..., 5$ that avoid sampling from the full joint posterior and computing the GP on the stochastic locations.

Time complexity: $\mathcal{O}(K^3)$ **Space complexity:** $\mathcal{O}(K^2)$

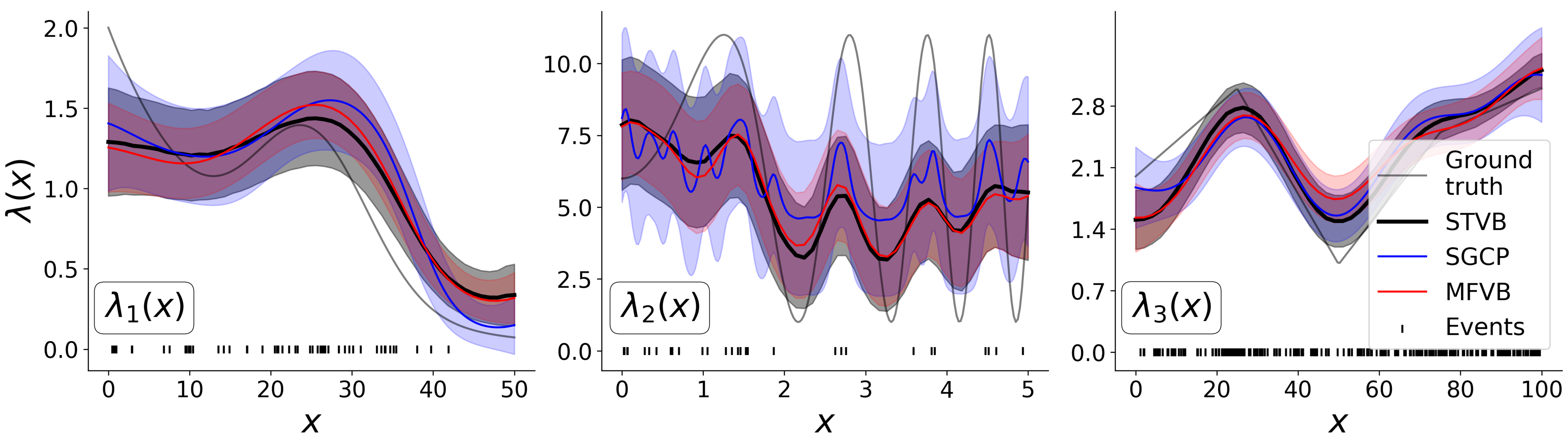
Key References

[1] Adams, R. P., Murray, I., and MacKay, D. J. *Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities* Proceedings of the 26th Annual International Conference on Machine Learning, pages 9–16 (2009).

[2] Lloyd, C., Gunter, T., Osborne, M. A., and Roberts, S. J. *Variational Inference for Gaussian Process Modulated Poisson Processes* International Conference on Machine Learning, pages 1814-1822 (2015).

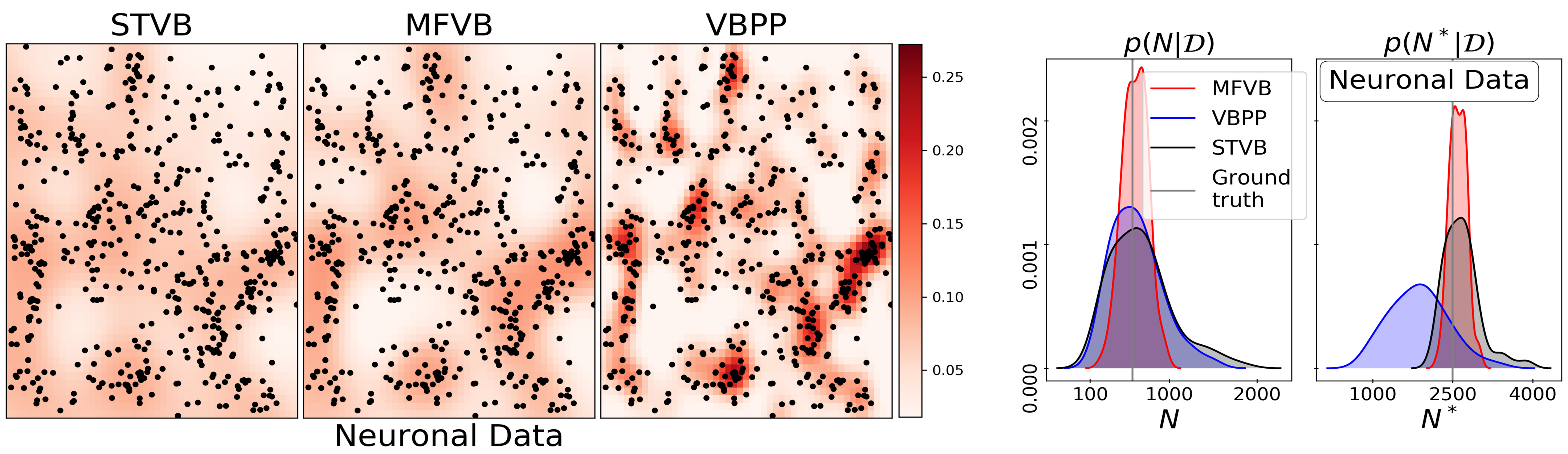
[3] Donner, C. and Opper, M. *Efficient bayesian inference of sigmoidal gaussian cox processes* The Journal of Machine Learning Research, 9(1):2710–2743 (2018).

Synthetic data



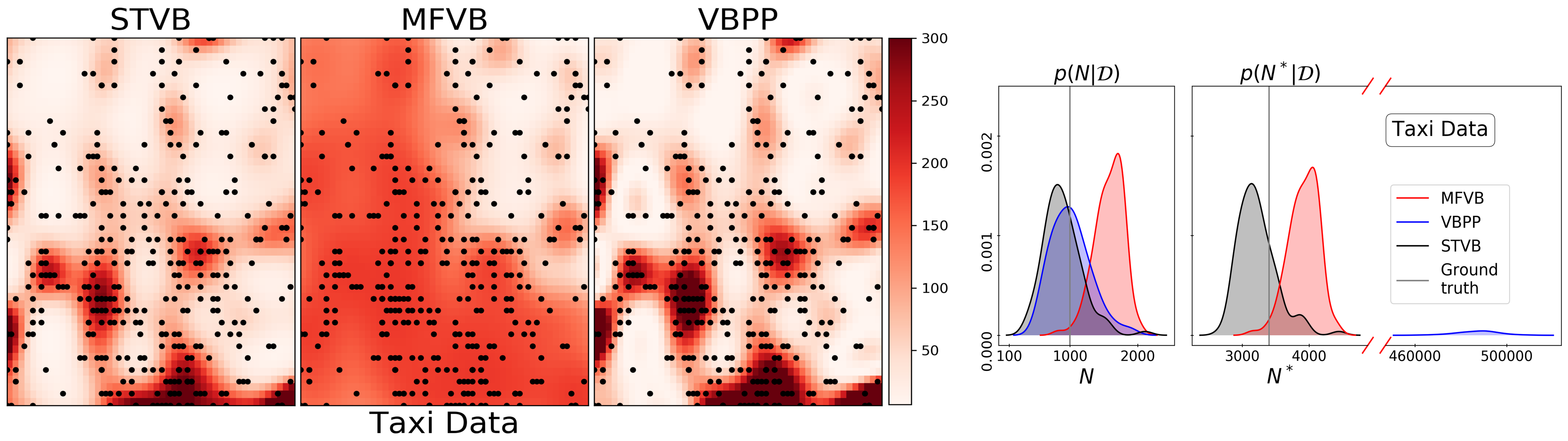
	$\lambda_1(x)$ l_2	ℓ_{test}	NLPL	EC 30% CI	EC 40% CI	$\lambda_2(x)$ l_2	ℓ_{test}	NLPL	EC 30% CI	EC 40% CI	$\lambda_3(x)$ l_2	ℓ_{test}	NLPL	EC 30% CI	EC 40% CI	CPU time (s)
STVB	**3.44** (1.43)	**-1.39** (1.05)	4.71 (0.51)	**0.81** (0.27)	**0.72** (0.27)	46.28 (9.95)	56.04 (4.47)	5.62 (0.72)	**0.91** (0.24)	**0.88** (0.23)	**7.39** (2.76)	153.98 (11.91)	6.41 (0.64)	**0.99** (0.03)	0.97 (0.09)	315.59
MFVB	4.56 (1.43)	-2.84 (1.0)	4.74 (0.1)	0.76 (0.25)	0.61 (0.28)	44.44 (10.7)	55.35 (4.72)	5.52 (1.29)	0.89 (0.23)	0.84 (0.29)	8.17 (3.43)	155.08 (10.20)	5.82 (0.61)	0.97 (0.09)	0.91 (0.14)	0.01
VBPP	9.19 (2.32)	-7.71 (3.31)	8.91 (1.19)	0.75 (0.21)	0.41 (0.25)	48.15 (13.16)	**56.82** (4.42)	5.20 (1.33)	0.76 (0.26)	0.45 (0.26)	20.54 (6.53)	152.82 (11.43)	8.35 (2.28)	0.83 (0.19)	0.43 (0.14)	0.44
SGCP	4.22 (1.88)	**-1.39** (1.28)	**4.21** (1.04)	0.39 (0.28)	0.27 (0.22)	**43.50** (8.69)	55.05 (1.35)	**3.77** (0.54)	0.64 (0.09)	0.14 (0.05)	14.44 (2.97)	**165.66** (2.12)	**4.78** (0.33)	0.49 (0.03)	0.34 (0.07)	2764.88
LGCP	67.76 (24.38)	-5.26 (8.84)	26.26 (8.09)	0.08 (0.12)	0.03 (0.09)	106.74 (13.89)	28.56 (6.88)	15.75 (3.36)	0.04 (0.08)	0.00 (0.00)	19.24 (6.44)	147.67 (11.76)	10.84 (1.36)	**0.99** (0.00)	**0.99** (0.12)	4.74

Application I: Neuronal data



	Neuronal data				
	$\ell_{test}[\times 10^3]$	NLPL	EC-30% CI	EC-40% CI	CPU time (s)
STVB	-84.55 (16.05)	**10.10** (7.02)	**1.00-1.00** (0.00)-(0.00)	**0.99-0.56** (0.10)-(0.50)	193.07
MFVB	**-83.54** (4.60)	10.71 (3.39)	**1.00**-0.03 (0.00)-(0.17)	0.78-0.00 (0.41)-(0.00)	0.35
VBPP	-83.89 (12.49)	11.39 (8.18)	**1.00**-0.00 (0.00) - (0.00)	0.83-0.00 (0.38)-(0.00)	26.23

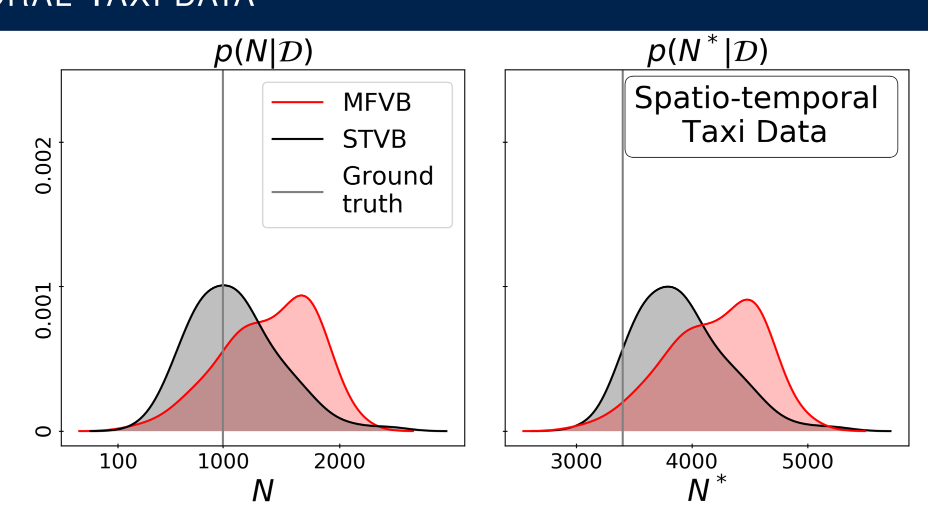
Application II: Taxi data



	Taxi data				
	$\ell_{test}[\times 10^3]$	NLPL	EC-30% CI	EC-40% CI	CPU time (s)
STVB	**-27.96** (9.16)	**27.96** (9.16)	0.81-**0.37** (0.39)-(0.48)	0.09-**0.01** (0.29)-(0.10)	290.34
MFVB	-40.8 (6.41)	40.65 (6.41)	0.00-0.00 (0.00)-(0.00)	0.00-0.00 (0.00)-(0.00)	0.24
VBPP	-31.32 (8.18)	31.32 (8.18)	**0.98**-0.00 (0.14)-(0.00)	**0.48**-0.00 (0.50)-(0.00)	3.62

Application III: Spatio-temporal Taxi data

	Spatio-temporal Taxi Data				
	$\ell_{test}[\times 10^7]$	NLPL$[\times 10^5]$	EC-30% CI	EC-40% CI	CPU time (s)
STVB	**-31.26** (10.88)	**31.26** (10.88)	**1.00**-0.00 (0.00)-(0.00)	**0.98**-0.00 (0.14)-(0.00)	1208.00
VBPP	-42.97 (9.56)	42.97 (9.56)	0.00-0.00 (0.00)-(0.00)	0.00-0.00 (0.00)-(0.00)	1.00



Future Research

- Test the algorithm in higher dimensional settings.
- Develop a scalable fully structured variational inference scheme by relaxing the factorization assumption in the posterior.

More information

Paper: https://arxiv.org/pdf/1906.03161.pdf **Python Code:** https://github.com/VirgiAgl/STVB