Sparse Gaussian Processes Revisited: Bayesian Approaches to Inducing-Variable Approximations

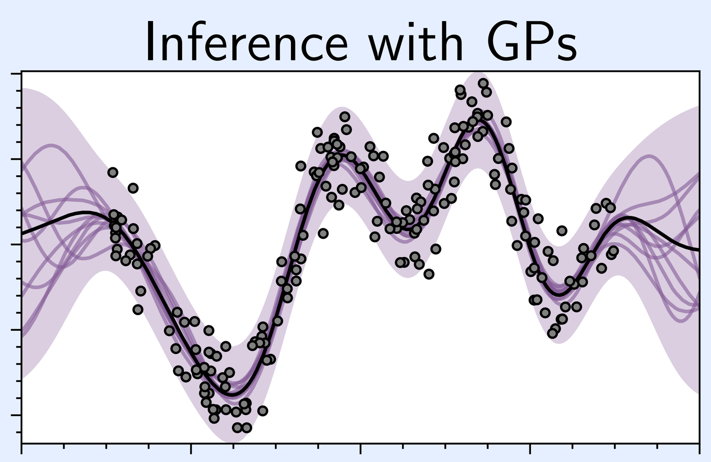
Simone Rossi[1], Markus Heinonen[2], Edwin V. Bonilla[3], Zheyang Shen[2], Maurizio Filippone[1]

[1] EURECOM (France), [2] Aalto University (Finland) , [3] CSIRO's Data61 (Australia)

Summary and Contributions

Gaussian processes (GPs) are a family of powerful non-parametric models, but computationally intractable.

For example,



Prior: $p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}))$

Likelihood: $p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2 \mathbf{I})$

Predictive posterior and marginal:

$$p(\mathbf{f}_\star \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{K}_\star(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y}, \mathbf{K}_{\star\star} - \mathbf{K}_\star(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{K}_\star^\top\right)$$

$$\log p(\mathbf{y}) = -\tfrac{1}{2}\mathbf{y}^\top(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y} - \tfrac{1}{2}\log\det(\mathbf{K}+\sigma^2\mathbf{I}) + \text{const.}$$

Sparse GPs: Sparse approximation methods introduce M *inducing variables* $\mathbf{u} = (u_1, \dots, u_M)$ drawn from the same prior at *inducing inputs* $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_M\}$ and they are solved for the augmented model $p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})$.

Contributions:

- ▶ We treat $\mathbf{Z}$, $\mathbf{u}$ and $\boldsymbol{\theta}$ in a Bayesian way by sampling from the true posterior via (SG)HMC in both shallow and deep GPs
- ▶ We analyze a number of priors on $\mathbf{Z}$ and with this setup we revisit sparse approximations prior to Titsias [4]

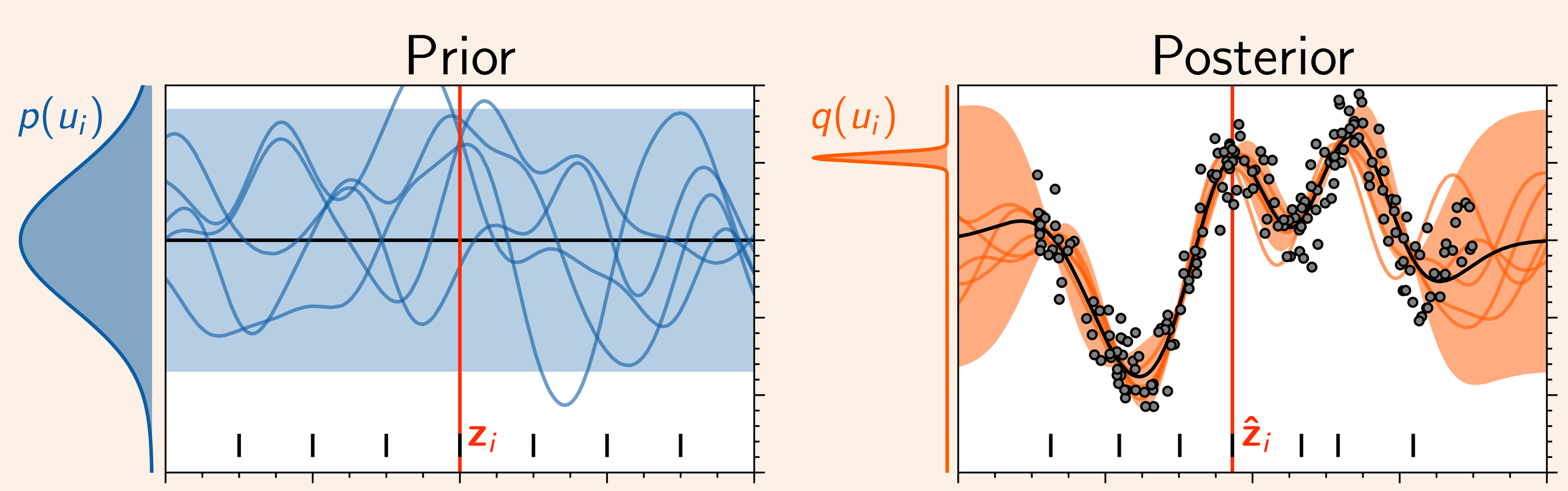
The classic recipe for inference of sparse GPs

SVGP [2] uses variational inference to approximate the posterior:

1. Factorize the posterior $q(\mathbf{f}, \mathbf{u})$ as $q(\mathbf{u})p(\mathbf{f} \mid \mathbf{u})$
2. Define approximate posterior for $\mathbf{u}$, e.g. $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$
3. Compute a lower bound to the marginal likelihood

$$\mathcal{L}_{\text{ELBO}} \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{u})p(\mathbf{f} \mid \mathbf{u})} \log p(\mathbf{y} \mid \mathbf{f}) - \text{KL}\left[q(\mathbf{u}) \parallel p(\mathbf{u})\right]$$

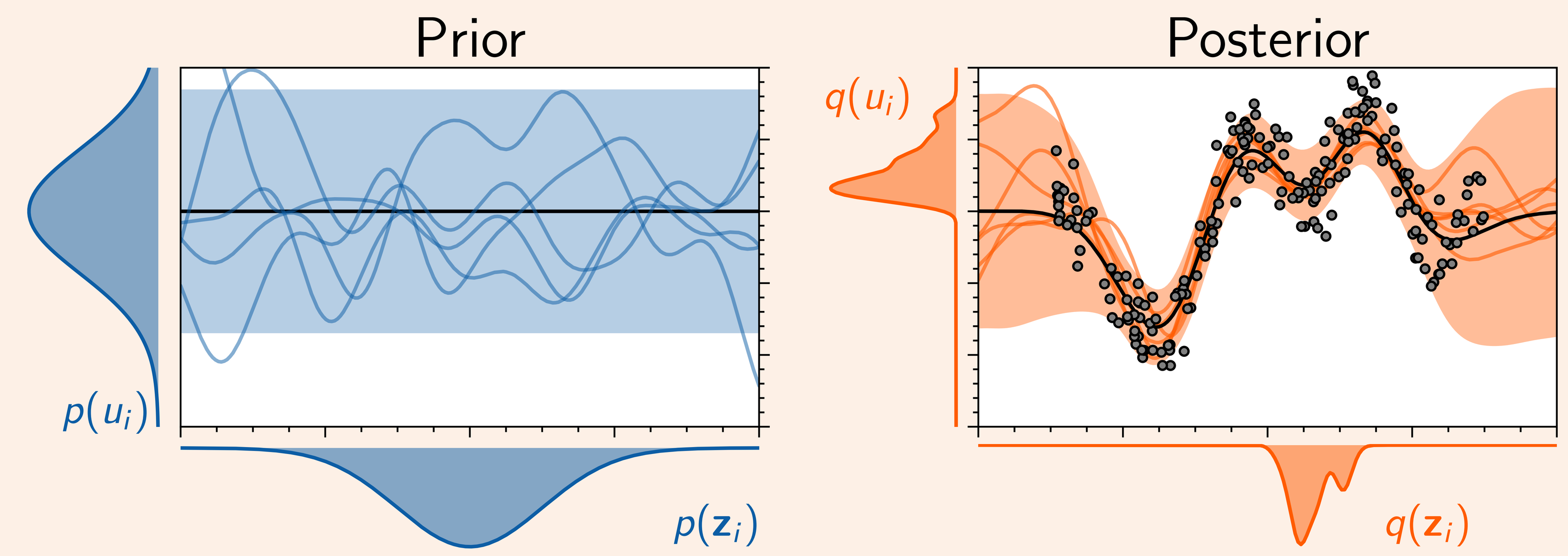
4. Maximize the lower bound w.r.t variational parameters (e.g. $\mathbf{m}, \mathbf{S}$), the covariance parameters $\boldsymbol{\theta}$ and $\mathbf{Z}$ jointly.



BSGP: Bayesian Sparse Gaussian Process

Consider the general formulation of the joint distribution,

$$p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{u}, \mathbf{f}, \mathbf{y}) = p(\boldsymbol{\theta})p(\mathbf{Z})p(\mathbf{u} \mid \mathbf{Z}, \boldsymbol{\theta})p(\mathbf{f} \mid \mathbf{u}, \mathbf{Z}, \boldsymbol{\theta})p(\mathbf{y} \mid \mathbf{f})$$



For (SG)HMC, we need objectives that factorize over observations. Two ways to proceed:

1. **VFE argument**, by minimization of the KL divergence

$$\mathcal{L}_{\text{VFE}} \stackrel{\text{def}}{=} \sum_n \mathbb{E}_{p(f_n \mid \boldsymbol{\theta}, \mathbf{Z}, \mathbf{u})} \log p(y_n \mid f_n) + \log p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{u}) \quad \text{[see 3]}$$

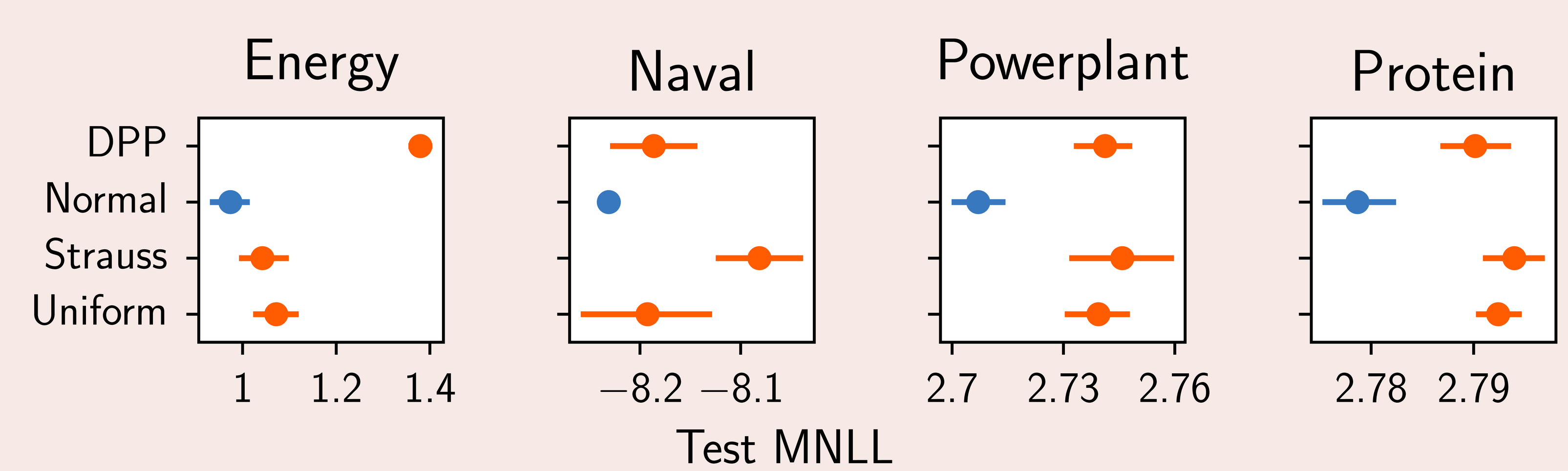
2. **FITC argument**, by independence of the conditional $\mathbf{f} \mid \mathbf{u}, \mathbf{Z}, \boldsymbol{\theta}$

$$\mathcal{L}_{\text{FITC}} \stackrel{\text{def}}{=} \sum_n \log \mathbb{E}_{p(f_n \mid \boldsymbol{\theta}, \mathbf{Z}, \mathbf{u})} p(y_n \mid f_n) + \log p(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{u}) \quad \text{[see 4]}$$

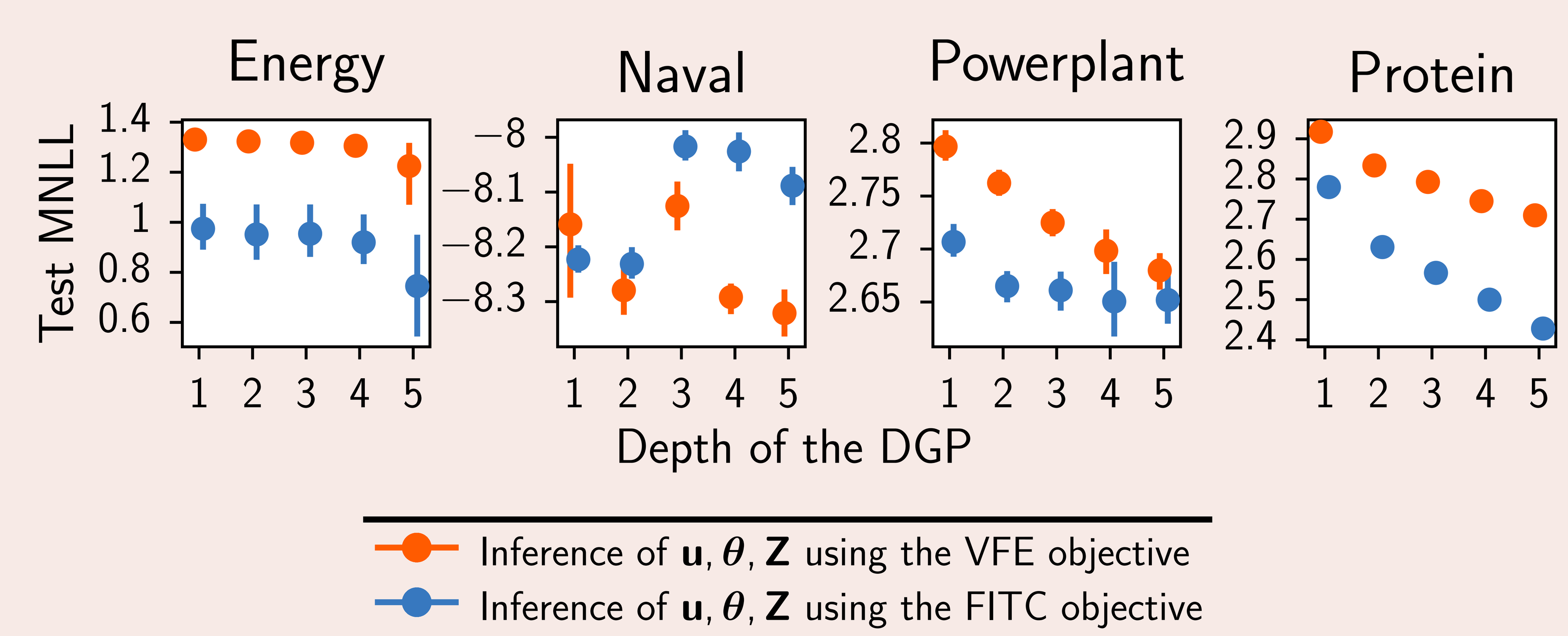
Which prior on Z should we choose?

DPP: $p(\mathbf{Z}) \propto \det \mathbf{K}_{zz}(\boldsymbol{\theta})$ Normal: $p(\mathbf{Z}) = \prod_{j=1}^M \mathcal{N}(\mathbf{z}_j \mid \mathbf{0}, \mathbf{I})$

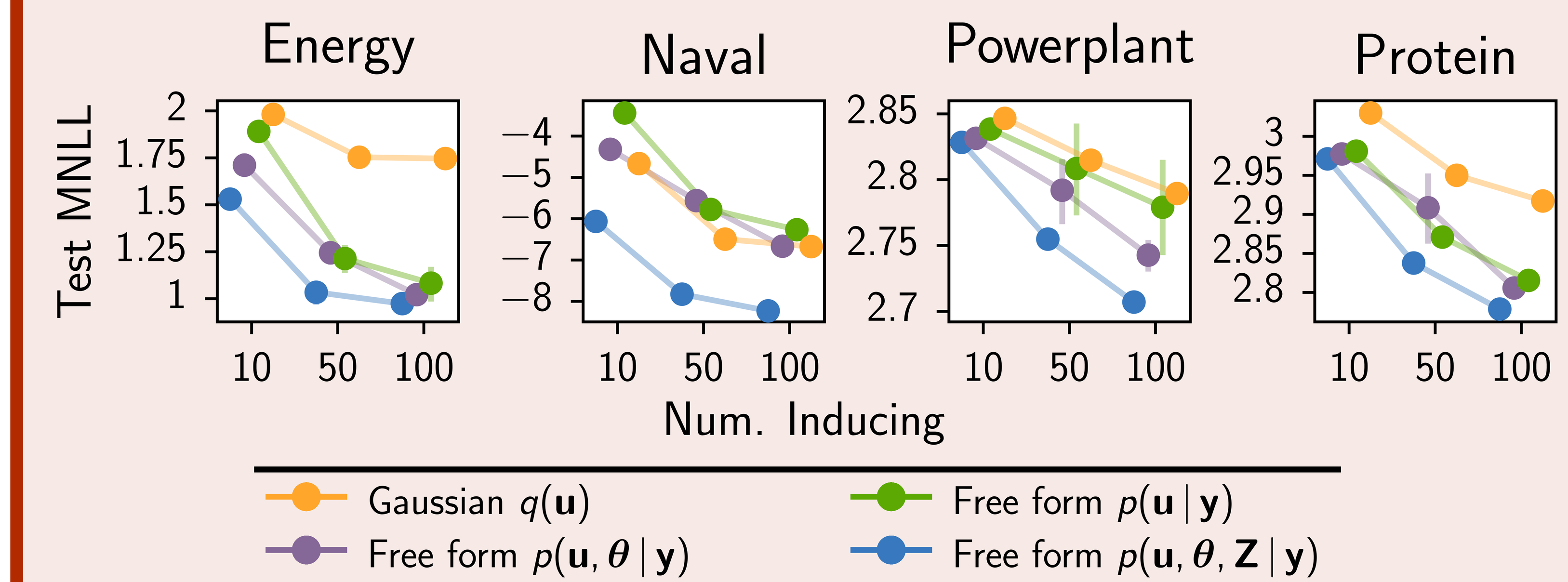
Strauss: $p(\mathbf{Z}) \propto \lambda^M \gamma^{\sum_{\mathbf{z},\mathbf{z}' \in \mathbf{Z}} \delta(\lvert \mathbf{z}-\mathbf{z}' \rvert < r)}$ Uniform: $p(\mathbf{Z}) = 0$

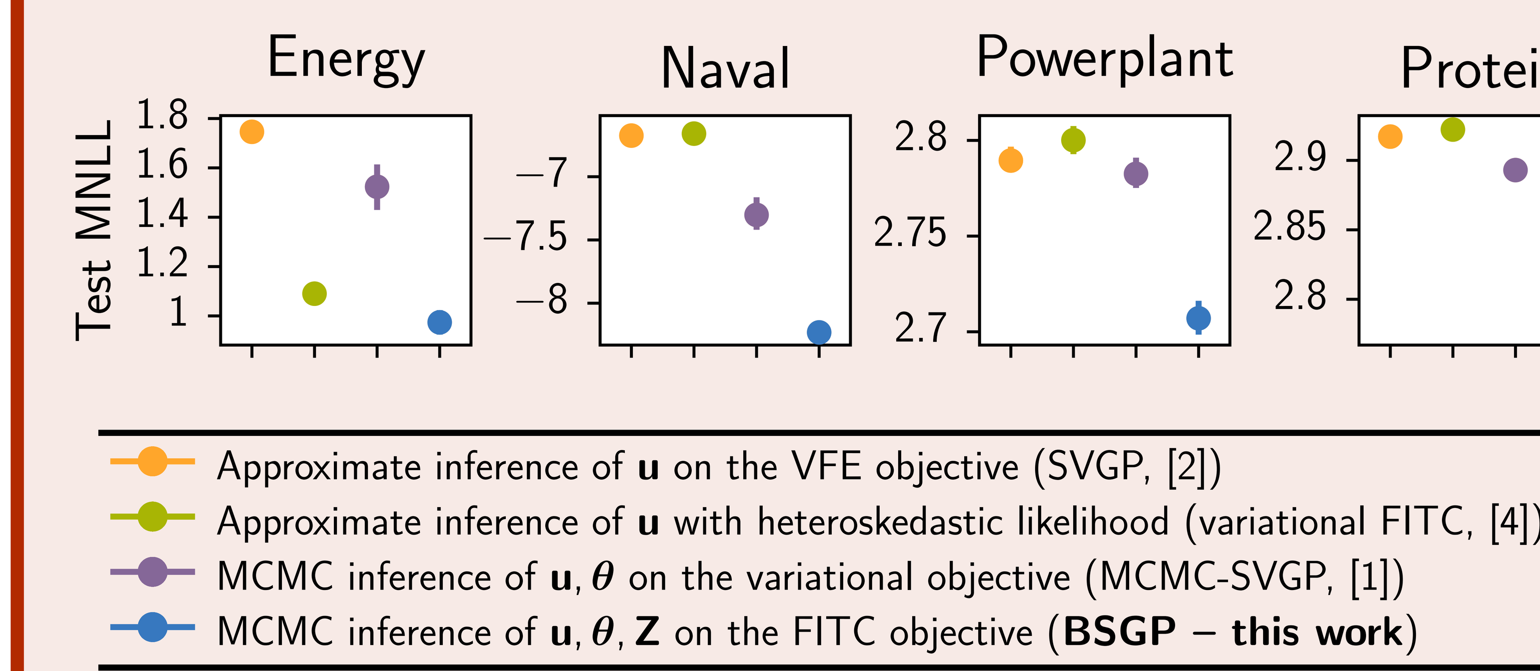


With VFE or FITC?

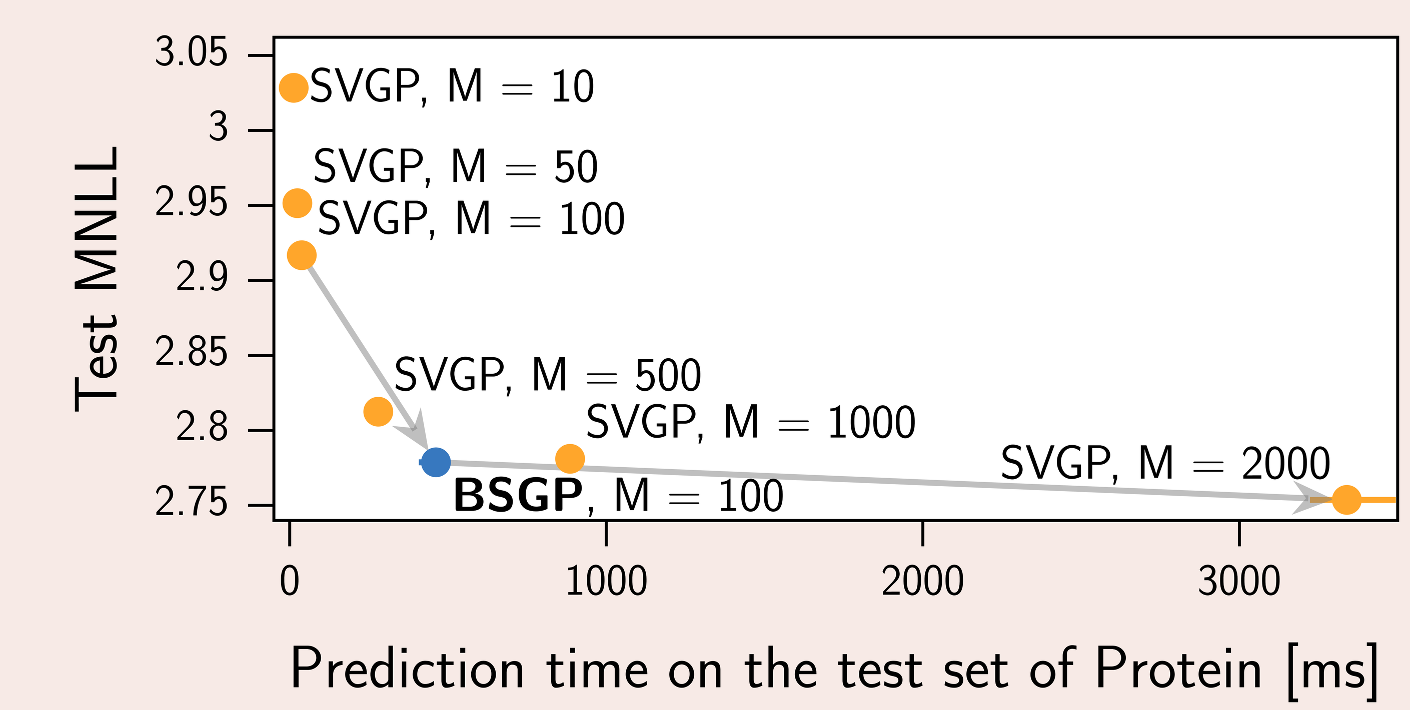


Does it help to be Bayesian about Z?





How expensive should we make SVGP to match BSGP?



References

[1] J. Hensman et al. "MCMC for Variationally Sparse Gaussian Processes". *NeurIPS* 2015.

[2] J. Hensman et al. "Scalable Variational Gaussian Process Classification". *AISTATS* 2015.

[3] E. Snelson and Z. Ghahramani. "Sparse Gaussian Processes using Pseudo-Inputs". *NeurIPS* 2006.

[4] M. K. Titsias. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". *AISTATS* 2009.