Improving the Gaussian Process Sparse Spectrum Approximation by Representing Uncertainty in Frequency Inputs

Yarin Gal • Richard Turner

yg279@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

The Gaussian process (GP)

- Is awesome
- ... but with a great computational cost – $\mathcal{O}(N^3)$ time complexity for N data points:

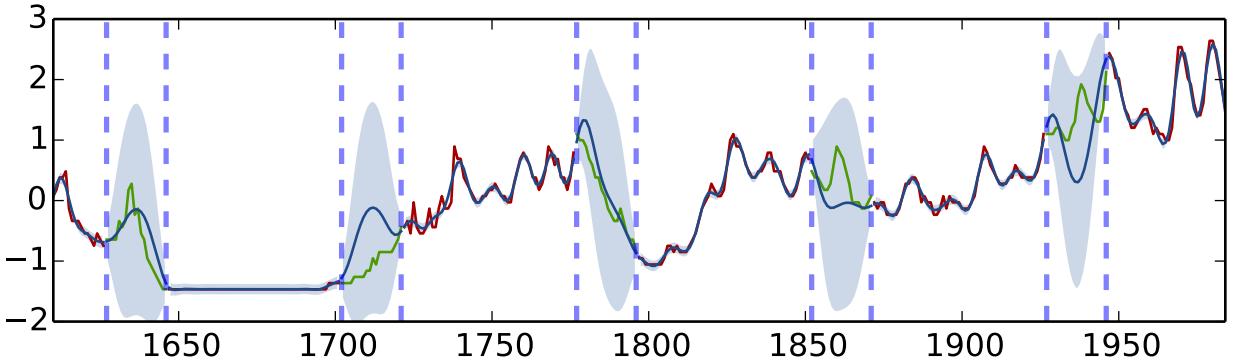
$$p(\mathbf{Y}|\mathbf{X}) = \mathcal{N}\left(\mathbf{Y}; \mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{X}) + \tau^{-1}\mathbf{I}_N\right)$$

with Q dimensional input $\mathbf{x}$, D dimensional output $\mathbf{y}$, and stationary covariance function $\mathbf{K}$.

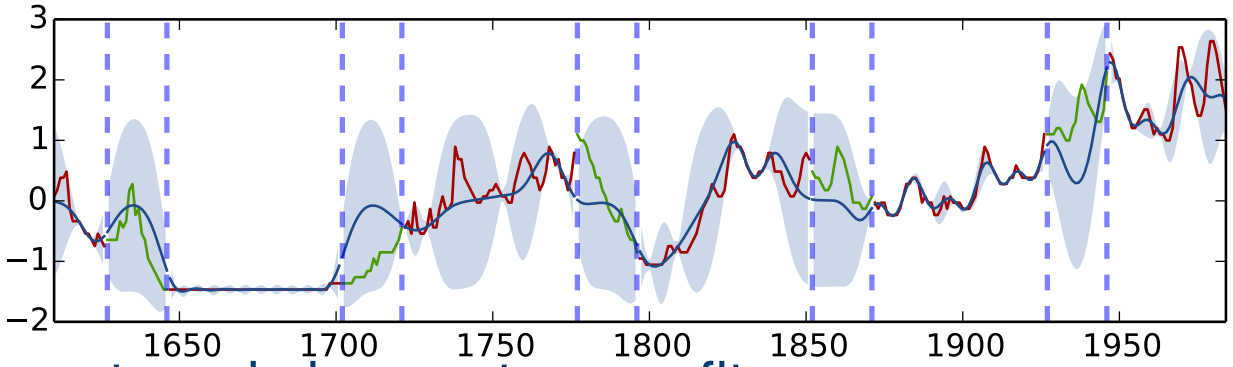
Many Approximations

UNIVERSITY OF CAMBRIDGE

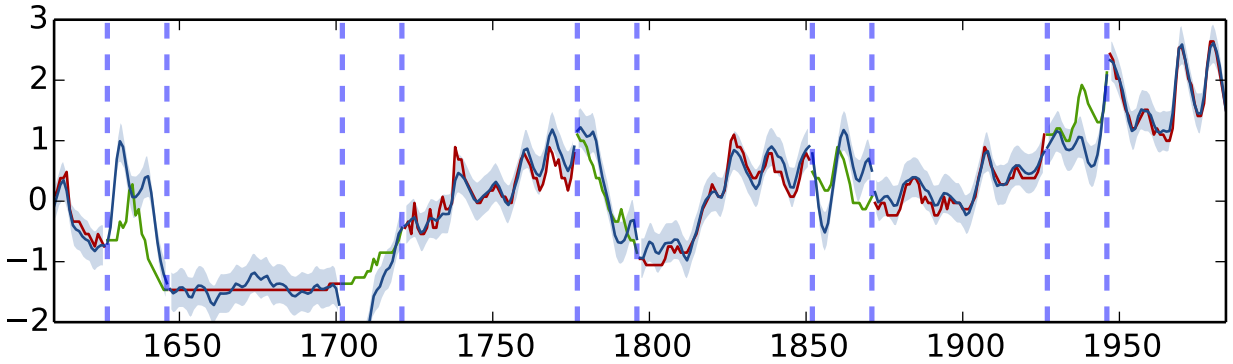
Full GP:



- Sparse pseudo-input cannot handle complex functions well:



- Sparse spectrum is known to over-fit:



Main Idea:

UNIVERSITY OF CAMBRIDGE

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour
- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a **random covariance function**
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference

Main Idea:

UNIVERSITY OF CAMBRIDGE

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour
- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a **random covariance function**
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference

Main Idea:

UNIVERSITY OF CAMBRIDGE

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour
- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a **random covariance function**
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference

Main Idea:

UNIVERSITY OF CAMBRIDGE

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour
- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a **random covariance function**
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference

Main Idea:

UNIVERSITY OF CAMBRIDGE

- **Variational Sparse Spectrum GP (VSSGP)**
  - use variational inference for the sparse spectrum approximation
  - avoids over-fitting, efficiently captures globally complex behaviour
- **In short—**
  - we replace the GP covariance function with a finite Monte Carlo approximation
  - we view this as a **random covariance function**
  - conditioned on data this random variable has an intractable posterior
  - we approximate this posterior with variational inference

In more detail (with a squared exponential covariance function)—

Given Fourier transform of the covariance function:

$$\begin{aligned}\mathbf{K}(\mathbf{x}-\mathbf{y}) &= \sigma^2 e^{-\frac{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}{2}} \\ &= \sigma^2 \int \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_Q) \cos\left(2\pi \mathbf{w}^T(\mathbf{x}-\mathbf{y})\right) d\mathbf{w}.\end{aligned}$$

Fourier transform of the squared exponential covariance function:

$$\mathbf{K}(\mathbf{x}-\mathbf{y}) = \sigma^2 \int \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_Q) \cos\left(2\pi \mathbf{w}^T(\mathbf{x}-\mathbf{y})\right) \mathrm{d}\mathbf{w},$$

Auxiliary variable b:

$$\mathbf{K}(\mathbf{x}-\mathbf{y}) = 2\sigma^2 \int \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_Q)\mathrm{Unif}[0, 2\pi] \cos\left(2\pi \mathbf{w}^T\mathbf{x} + b\right) \cos\left(2\pi \mathbf{w}^T\mathbf{y} + b\right) \mathrm{d}\mathbf{w}\mathrm{d}b.$$

Auxiliary variable b:

$$\mathbf{K}(\mathbf{x}-\mathbf{y}) = 2\sigma^2 \int \mathcal{N}(\mathbf{w}; 0, \mathbf{I}_Q)\text{Unif}[0, 2\pi] \cos\left(2\pi \mathbf{w}^T \mathbf{x} + b\right) \cos\left(2\pi \mathbf{w}^T \mathbf{y} + b\right) \text{d}\mathbf{w}\text{d}b,$$

Monte Carlo integration with K terms:

$$\widehat{\mathbf{K}}(\mathbf{x}-\mathbf{y}) = \frac{2\sigma^2}{K} \sum_{k=1}^{K} \cos\left(2\pi \mathbf{w}_k^T \mathbf{x} + b_k\right) \cos\left(2\pi \mathbf{w}_k^T \mathbf{y} + b_k\right)$$

with $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q)$, $b_k \sim \text{Unif}[0, 2\pi]$.

This is a **random covariance function**.

Monte Carlo integration with K terms:

$$\widehat{\mathbf{K}}(\mathbf{x}-\mathbf{y}) = \frac{2\sigma^2}{K}\sum_{k=1}^{K}\cos\left(2\pi\mathbf{w}_k^T\mathbf{x}+b_k\right)\cos\left(2\pi\mathbf{w}_k^T\mathbf{y}+b_k\right),$$

Rewrite the covariance function with $\Phi \in \mathbb{R}^{N\times K}$

$$\mathbf{w}_k \sim \mathcal{N}(0,\mathbf{I}_Q), \quad b_k \sim \text{Unif}[0,2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^{K}$$

$$\Phi_{n,k}(\omega) = \sqrt{\frac{2\sigma^2}{K}}\cos\left(2\pi\mathbf{w}_k^T\mathbf{x}_n+b_k\right),$$

$$\widehat{\mathbf{K}}(\mathbf{x}-\mathbf{y}) = \Phi(\omega)\Phi(\omega)^T.$$

Rewrite the covariance function with $\Phi \in \mathbb{R}^{N \times K}$

$$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^K$$

$$\Phi_{n,k}(\omega) = \sqrt{\frac{2\sigma^2}{K}} \cos\left(2\pi \mathbf{w}_k^T \mathbf{x}_n + b_k\right),$$

$$\widehat{\mathbf{K}}(\mathbf{x} - \mathbf{y}) = \Phi(\omega)\Phi(\omega)^T,$$

Integrate the GP over the random covariance function

$$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^K$$

$$p(\mathbf{Y}|\mathbf{X}, \omega) = \mathcal{N}\left(\mathbf{Y}; \mathbf{0}, \Phi(\omega)\Phi(\omega)^T + \tau^{-1}\mathbf{I}_N\right)$$

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \omega)\, p(\omega) \mathrm{d}\omega$$

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \omega) p(\omega|\mathbf{X}, \mathbf{Y}) \mathrm{d}\omega.$$

Integrate the GP over the random covariance function

$$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^K$$

$$p(\mathbf{Y}|\mathbf{X}, \omega) = \mathcal{N}\left(\mathbf{Y}; \mathbf{0}, \Phi(\omega)\Phi(\omega)^T + \tau^{-1}\mathbf{I}_N\right)$$

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \omega)p(\omega)\mathrm{d}\omega$$

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \omega)p(\omega|\mathbf{X}, \mathbf{Y})\mathrm{d}\omega.$$

Use variational distribution $q(\omega) = \prod q(\mathbf{w}_k)q(b_k)$ to approximate posterior $p(\omega|\mathbf{X}, \mathbf{Y})$:

$$q(\mathbf{w}_k) = \mathcal{N}(\mu_k, \Sigma_K), \quad q(b_k) = \text{Unif}(\alpha_k, \beta_k),$$

with Σ_K diagonal.

Maximise log evidence lower bound

$$\mathcal{L}_{VSSGP} = \frac{1}{2}\sum_{d=1}^{D}\Big(\log(|\tau^{-1}\,\mathbf{\Sigma}\,|) + \tau\mathbf{y}_d^T\,E_{q(\omega)}(\Phi)\,\mathbf{\Sigma}\,E_{q(\omega)}(\Phi^T)\,\mathbf{y}_d + \ldots\Big)$$
$$- \mathrm{KL}(q(\omega)||p(\omega))$$

with $\mathbf{\Sigma} = (E_{q(\omega)}(\Phi^T\Phi) + \tau^{-1}I)^{-1}$.

Maximise log evidence lower bound

$$\mathcal{L}_{VSSGP} = \frac{1}{2}\sum_{d=1}^{D}\Big(\log(|\tau^{-1}\,\mathbf{\Sigma}\,|) + \tau\mathbf{y}_d^T\,E_{q(\omega)}(\Phi)\,\mathbf{\Sigma}\,E_{q(\omega)}(\Phi^T)\,\mathbf{y}_d + \ldots\Big) - \mathrm{KL}(q(\omega)||p(\omega))$$

with $\mathbf{\Sigma} = (E_{q(\omega)}(\Phi^T\Phi) + \tau^{-1}I)^{-1}$.

KL and expectations analytical with

$$E_{q(\mathbf{w})}\big(\cos(\mathbf{w}^T\mathbf{x} + b)\big) = e^{-\frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}}\cos(\mu^T\mathbf{x} + b).$$

Requires $\mathcal{O}(NK^2 + K^3)$ **time complexity**.
Parallel inference with T workers: $\searrow\ \mathcal{O}(\frac{NK^2}{T} + K^3)$.

UNIVERSITY OF CAMBRIDGE

Factorised VSSGP (fVSSGP)

- We often use large K.
- K by K matrix inversion is still slow: $\mathcal{O}(K^3)$.
- **It is silly to invert the whole matrix every time**
  — slightly changing the parameters we expect the inverse to not change too much.
- We can do better with an additional **auxiliary variable**.

We integrated the GP over the random covariance function

$$\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{I}_Q), \quad b_k \sim \text{Unif}[0, 2\pi], \quad \omega = \{\mathbf{w}_k, b_k\}_{k=1}^K$$

$$p(\mathbf{Y}|\mathbf{X}, \omega) = \mathcal{N}\left(\mathbf{Y}; \mathbf{0}, \Phi(\omega)\Phi(\omega)^T + \tau^{-1}\mathbf{I}_N\right)$$

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \omega)p(\omega)\mathrm{d}\omega,$$

Introduce auxiliary random variables $\mathbf{A} \in \mathbb{R}^{K \times D}$

$$\mathbf{A} \sim \mathcal{N}(0, \mathbf{I}_{K \times D}),$$

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{A}, \omega) = \mathcal{N}\left(\mathbf{Y}; \Phi(\omega)\mathbf{A}, \tau^{-1}\mathbf{I}_N\right)$$

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, A, \omega)p(A)p(\omega)\mathrm{d}\omega\mathrm{d}\mathbf{A}.$$

Introduce auxiliary random variables $\mathbf{A} \in \mathbb{R}^{K \times D}$

$$\mathbf{A} \sim \mathcal{N}(0, \mathbf{I}_{K \times D}),$$
$$p(\mathbf{Y}|\mathbf{X}, \mathbf{A}, \omega) = \mathcal{N}\left(\mathbf{Y}; \Phi(\omega)\mathbf{A}, \tau^{-1}\mathbf{I}_N\right)$$
$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, A, \omega)p(A)p(\omega)\mathrm{d}\omega\mathrm{d}\mathbf{A},$$

Use variational distribution $q(\omega) = \prod q(\mathbf{w}_k)q(b_k) \prod q(\mathbf{a}_d)$ to approximate posterior $p(\omega, \mathbf{A}|\mathbf{X}, \mathbf{Y})$:

$$q(\mathbf{a}_d) = \mathcal{N}(\mathbf{m}_d, \mathbf{s}_d)$$

over the rows of A with $\mathbf{s}_d$ diagonal.

Maximise log evidence lower bound

$$\mathcal{L}_{fVSSGP} = \sum_{d=1}^{D} \Big(\tau \mathbf{y}_d^T \, E_{q(\omega)}(\Phi)\mathbf{m}_d \; - \frac{\tau}{2}\mathrm{tr}\big(\, E_{q(\omega)}(\Phi^T\Phi)(\mathbf{s}_d + \mathbf{m}_d\mathbf{m}_d^T) \, \big) + \ldots \Big) - \mathrm{KL}(q(\mathbf{A})||p(\mathbf{A})) - \mathrm{KL}(q(\omega)||p(\omega)).$$

Requires $\mathcal{O}(NK^2)$ **time complexity** — no matrix inversion.
Parallel inference with T workers $\searrow \mathcal{O}(\frac{NK^2}{T})$.

Stochastic Factorised VSSGP (sfVSSGP)

- We often use large N.
- N matrix products of size $K \times K$ is still slow: $\mathcal{O}(NK^2)$.
- **It is silly to evaluate the objective over the entire dataset** — might have redundant data.
- We can do even better with **stochastic optimisation**.

Maximise log evidence lower bound

$$\mathcal{L}_{sfVSSGP} \approx \frac{N}{|S|} \sum_{n \in S} \sum_{d=1}^{D} \mathcal{L}_{nd} - \mathrm{KL}(q(\mathbf{A})||p(\mathbf{A})) - \mathrm{KL}(q(\omega)||p(\omega))$$

with random data subset S this is an unbiased estimator of $\mathcal{L}_{fVSSGP}$.

$\mathcal{O}(SK^2)$ **time complexity** with $S << N$ size of random subset.

Time Complexity Comparison

UNIVERSITY OF CAMBRIDGE

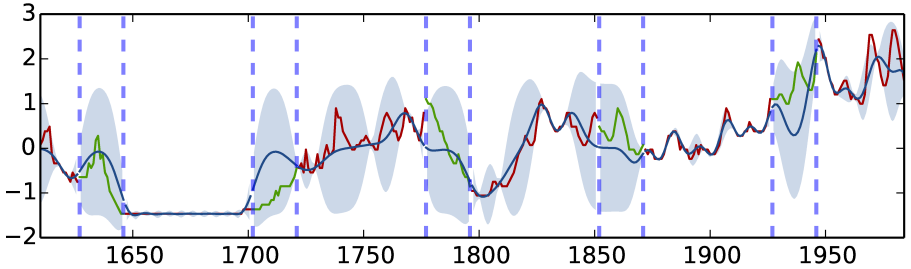
Full GP	$\mathcal{O}(N^3)$
SPGP / SSGP / **VSSGP**	$\mathcal{O}(NK^2 + K^3)$
Factorised SPGP	$\mathcal{O}(NK^2 + K^3)$
Factorised VSSGP	$\mathcal{O}(NK^2)$
Stochastic SPGP	$\mathcal{O}(SK^2 + K^3)$, $S << N$
Stochastic fVSSGP	$\mathcal{O}(SK^2)$, $S << N$

with K number of inducing points.

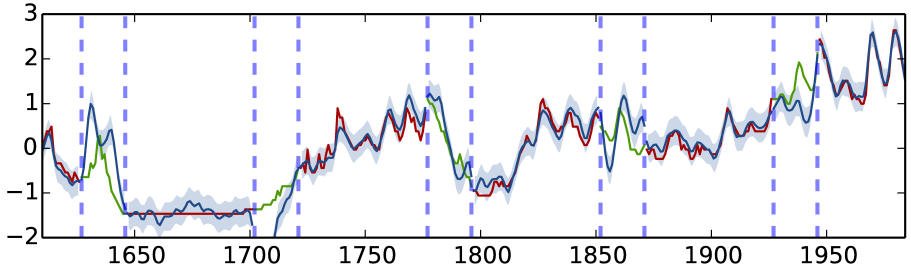
Example results

UNIVERSITY OF CAMBRIDGE

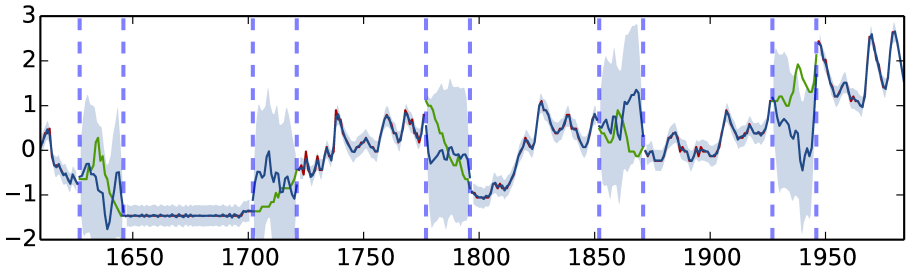
Interpolation on the reconstructed solar irradiance dataset (SE covariance function, $K = 50$ inducing inputs):



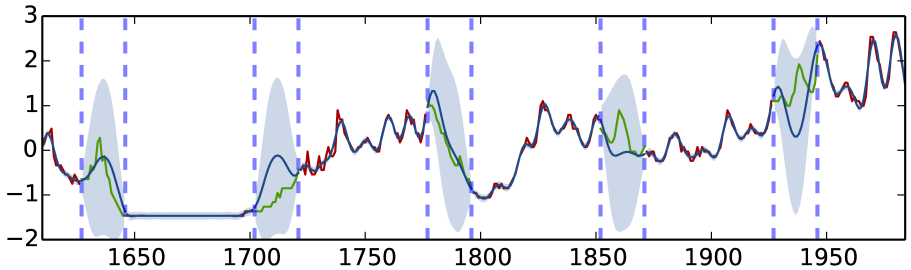
Sparse pseudo-input GP



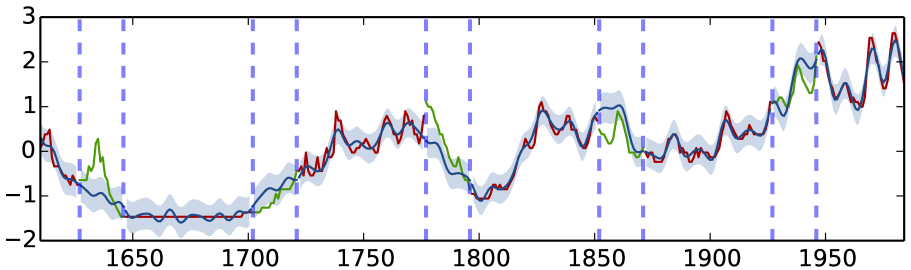
Sparse Spectrum GP



Random Projections ($K = 500$)



Full GP



Variational Sparse Spectrum GP

Solar	SPGP	SSGP	RP_1	RP_2	GP	**VSSGP**
Train	0.23	0.15	0.32	0.04	0.08	**0.13**
Test	0.61	0.63	0.65	0.76	0.50	**0.41**

Interpolation RMSE on train / test sets

Many more results

UNIVERSITY OF CAMBRIDGE

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger K)

Many more results

UNIVERSITY OF CAMBRIDGE

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger K)

Many more results

UNIVERSITY OF CAMBRIDGE

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger K)

Many more results

UNIVERSITY OF CAMBRIDGE

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger K)

Many more results

UNIVERSITY OF CAMBRIDGE

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger K)

Many more results

UNIVERSITY OF CAMBRIDGE

Many more results in the paper:

- Variational SSGP properties (uncertainty increases far from data)
- From SSGP to variational SSGP (big improvement)
- VSSGP, factorised VSSGP, and stochastic factorised VSSGP (more assumptions = worse results; still better than SSGP)
- Stochastic variational inference comparison (better than SPGP SVI)
- Speed-accuracy trade-offs (better accuracy with larger K)

Thank you