Variational Gaussian Processes For Linear Inverse Problems

Joint work with Botond Szabo

Thibault Randrianarisoa

2025 Workshop GAP ANR

July 9, 2025

Inverse problems

Statistical inverse regression models

Model Consider observations (X_i, Y_i) arising from

$$Y_i = \mathcal{A}f_0(X_i) + \epsilon_i, \qquad i = 1, \dots, n,$$

- $\mathcal{A}\colon L^2(\mathcal{T}, \mu) \mapsto L^2(\mathcal{X}, \nu)$ *known* injective continous linear operator
- $X_1, \dots, X_n \sim \nu$ iid
- $\epsilon_1, \dots, \epsilon_n \sim \mathcal{N}(0, 1)$ iid, independent of X_i's

Ex: Volterra operator, Radon transform, heat equation, deconvolution...

Goal Estimate f_0 in $L^2(\mu)$-loss $\ell(f, g) = \int_{\mathcal{T}} (f - g)^2(t) d\mu(t)$

Twofold challenge

- Inverting $\mathcal{A}$ (inverse problem theory)
- Denoising observations (statistics)

Ill-posedness

Non-continuous inverse

No *naive* plug-in estimator

Need for regularization in statistical approaches

e.g. Tikhonov regularization

$$\min_f \sum_i \left(\mathcal{A}f(X_i) - Y_i\right)^2 + \gamma \left\|f\right\|^2$$

Ill-posedness

Non-continuous inverse

No *naive* plug-in estimator

Need for regularization in statistical approaches

e.g. Tikhonov regularization

$$\min_f \sum_i \left(\mathcal{A}f(X_i) - Y_i\right)^2 + \gamma \left\|f\right\|^2$$

Other choice: Bayesian procedures

MAP with GP prior

$$\min_f \sum_i \left(\mathcal{A}f(X_i) - Y_i\right)^2 + \sigma^2 \left\|f\right\|_{\mathbb{H}}^2$$

Bayesian approaches

Posterior contraction rates

Bayesian setting

Given: data $(x_i, Y_i)_{i=1}^n$ **Infer**: posterior $\Pi[\cdot | X]$ from prior Π

Frequentist analysis of Bayesian procedures:

- Assume there exists f_0 such that $X \sim P_{f_0}$
- study the behaviour of $\Pi[\cdot | X]$ under P_{f_0}:
  - convergence to f_0
  - rate of convergence

$$E_{f_0}\Pi\left[f : \|f - f_0\| \geq M_n\varepsilon_n | X\right] \to 0, \quad M_n \to \infty, \tag{1}$$

Bayesian linear inverse problems

Growing interest in asymptotics of Bayesian approaches in last decade

- conjugate priors in mildly ill-posed problems [Knapik et al '14, Agapiou et al '13, Florens and Simoni '12]
- severely ill-posed problems [Knapik et al '14, Agapiou et al '14], e.g. initial condition heat equation
- rate adaptive Bayesian procedure [Knapik '13 & '16]
- non-conjugate priors [Ray '15]
- Uncertainty quantification [Szabó et al '15]
- General aproach to Bayesian inversion [Knapik and Salomond '18]

SVD

Assuming $\mathcal{A}^*\mathcal{A}$ compact,

$$\mathcal{A}^*\mathcal{A}f = \sum_l {\kappa_l}^2 \langle f, e_l \rangle e_l$$

Second basis f_l of $L^2(\mathcal{X})$ given by $\mathcal{A}e_l = \kappa_l f_l$

- mildly ill-posed: $\kappa_l \asymp l^{-p}$
- severely ill-posed: $\kappa_l \asymp e^{-cl^p}$

Diagonalized operator Easier to work in spectral domain

Gaussian processes

Gaussian processes are popular methods for inverse problems

Given SVD Centered GP $W = \sum_l \sqrt{\lambda_l}\, Z_l\, e_l$ with covariance kernel

$$K(x, y) = E(W_s W_t) = \sum_l \lambda_l e_l(s) e_l(t)$$

Gaussian processes

Gaussian processes are popular methods for inverse problems

Given SVD Centered GP $W = \sum_l \sqrt{\lambda_l}\, Z_l\, e_l$ with covariance kernel

$$K(x, y) = E(W_s W_t) = \sum_l \lambda_l e_l(s) e_l(t)$$

Covariance operator

$$\Lambda f(t) = \int K(s, t) f(s) d\mu(s)$$

is of trace class $\sum_l \lambda_l < \infty$

Gaussian processes

Gaussian processes are popular methods for inverse problems

Given SVD Centered GP $W = \sum_l \sqrt{\lambda_l}\, Z_l\, e_l$ with covariance kernel

$$K(x, y) = E(W_s W_t) = \sum_l \lambda_l e_l(s) e_l(t)$$

Covariance operator

$$\Lambda f(t) = \int K(s, t) f(s) d\mu(s)$$

is of trace class $\sum_l \lambda_l < \infty$

Covariance operator of $\mathcal{A}W$ also has discrete spectrum:

$$\mathcal{A}\Lambda\mathcal{A}^* f_l = \lambda_l \kappa_l^2 f_l$$

Sobolev class

Difficulty of estimation is measured by the minimax risk over some regularity class

$$\bar{H}^\beta := \left\{ f \in L^2(T;\mu) : \|f\|_\beta < \infty \right\}, \quad \|f\|_\beta^2 = \sum_j j^{2\beta} \left|\langle f, e_j \rangle\right|^2 ,$$

Minimax rate r_n^* is

- $\asymp n^{-\beta/(1+2\beta+2p)}$ if mildly ill-posed
- $\asymp \log^{-\beta/p} n$ if severely

GP Concentration result for inverse regression

Theorem (Posterior contraction)

For $f_0 \in \bar{H}^\beta$ and

1. *Mildly:* $\lambda_i \asymp i^{-1-2\beta}$, *for* β *large*
2. *Severely:* $\lambda_i \asymp i^{-\alpha} e^{-\xi i^p}$

There exists an event A_n, $P_0(A_n) \to 1$, *such that*

$$E_{f_0} \Pi \left[f : \ \|f - f_0\|_{L^2(T;\mu)} \geq M_n r_n^* \middle| \ X \right] \mathbb{1}_{A_n} \leq C e^{-cn(M_n r_n^*)^2}, \quad M_n \to \infty.$$

So, the SVD-related GP prior

- attains minimax rate if properly tuned
- works in both mildly and severely ill-posed settings

Sparse variational GPs

Time complexity

Computational drawback

Posterior is the GP

$$\mathcal{GP}\left(K_{\cdot n}\left(K_{nn}+\sigma^2 I_n\right)^{-1}\boldsymbol{Y},\ K(s,t)-K_{sn}\left(K_{nn}+\sigma^2 I_n\right)^{-1}K_{nt}\right)$$

- $K_{nn} = E_{\Pi}\boldsymbol{\mathcal{A}f}\boldsymbol{\mathcal{A}f}^T$ is the prior covariance at design points
- $K_{nt} = E_{\Pi}\boldsymbol{\mathcal{A}f}\mathcal{A}f(t)$

Time complexity

Computational drawback

Posterior is the GP

$$\mathcal{GP}\left(K_{\cdot n}\left(K_{nn}+\sigma^2 I_n\right)^{-1}\boldsymbol{Y},\ K(s,t)-K_{sn}\left(K_{nn}+\sigma^2 I_n\right)^{-1}K_{nt}\right)$$

- $K_{nn} = E_{\Pi}\boldsymbol{\mathcal{A}f}\boldsymbol{\mathcal{A}f}^T$ is the prior covariance at design points
- $K_{nt} = E_{\Pi}\boldsymbol{\mathcal{A}f}\mathcal{A}f(t)$

Issue: matrix inversion scales as $\mathcal{O}\left(n^3\right)$ in time

Time complexity

Computational drawback

Posterior is the GP

$$\mathcal{GP}\left(K_{\cdot n}\left(K_{nn}+\sigma^2 I_n\right)^{-1}\boldsymbol{Y},\; K(s,t)-K_{sn}\left(K_{nn}+\sigma^2 I_n\right)^{-1}K_{nt}\right)$$

- $K_{nn} = E_{\Pi}\boldsymbol{\mathcal{A}f}\boldsymbol{\mathcal{A}f}^T$ is the prior covariance at design points
- $K_{nt} = E_{\Pi}\boldsymbol{\mathcal{A}f}\mathcal{A}f(t)$

Issue: matrix inversion scales as $\mathcal{O}\left(n^3\right)$ in time

Solution: Low-rank approximation of K_{nn} [Seeger et al '03, Snelson and Ghahramani '05, Quiñonero Candela and Rasmussen '05, Titsias '09] to scale as $\mathcal{O}\left(nq^2\right)$

Variational approach of [Titsias '09]

Variational posterior. [Titsias '09] proposes to find the minimizer of KL divergence between posterior and

$$\mathcal{GP}\left(K_{\cdot q}K_{qq}^{-1}\boldsymbol{\mu},\ K(s,t) - K_{sq}K_{qq}^{-1}\left(K_{qq} - \Sigma\right)K_{qq}^{-1}K_{qt}\right)$$

- q inducing variables $u_1, \ldots, u_q$, i.e. point evaluations of the GP prior or continuous linear functionals of it
- $K_{qq}, K_{\cdot q}$ prior covariance of inducing variables
- $\boldsymbol{\mu}, \Sigma$ variational parameters, as we assume $\mathbf{u} \sim \mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right)$

Inducing variables

At minimum, KL is

$$\frac{1}{2}\left(\boldsymbol{Y}\left(\boldsymbol{Q}^{-1}-\boldsymbol{K}^{-1}\right)\boldsymbol{Y}+\log\frac{|\boldsymbol{Q}|}{|\boldsymbol{K}|}+\sigma^{-2}tr\left(\boldsymbol{K}-\boldsymbol{Q}\right)\right)$$

where $\boldsymbol{Q} = K_{nq}K_{qq}^{-1}K_{qn} + \sigma^2 I_n$ and $\boldsymbol{K} = K_{nn} + \sigma^2 I_n$ [rank-q approximation]

Inducing variables

At minimum, KL is

$$\frac{1}{2}\left(\boldsymbol{Y}\left(\boldsymbol{Q}^{-1}-\boldsymbol{K}^{-1}\right)\boldsymbol{Y}+\log\frac{|\boldsymbol{Q}|}{|\boldsymbol{K}|}+\sigma^{-2}tr\left(\boldsymbol{K}-\boldsymbol{Q}\right)\right)$$

where $\boldsymbol{Q}=K_{nq}K_{qq}^{-1}K_{qn}+\sigma^2 I_n$ and $\boldsymbol{K}=K_{nn}+\sigma^2 I_n$ [rank-q approximation]

Depends on the choice of q and $\mathbf{u}$!

Question: How large should q be ?

Inducing variables

This problem is related to the spectrum of K_{nn}, itself linked to the spectrum of the covariance operator $\mathcal{A}^* \Lambda \mathcal{A}$

Two choices [Burt et al '19]

- Eigendecomposition of covariance matrix: $(v_j^1, \ldots, v_j^n)$ jth eigenvector of K_{nn}

$$u_j = \sum_{i=1}^{n} v_j^i A f(x_i)$$

- Eigendecomposition of covariance operator:

$$u_j = \int_{\mathcal{X}} AW(x) f_j(x) dG(x)$$

Variational posterior contraction

Posterior results $E_{f_0}\Pi\left[f \in \mathcal{F}_n \middle| X\right]\mathbb{1}_{A_n} \leq Ce^{-\delta_n}$ gives

$$E_{f_0}\Psi\left[f \in \mathcal{F}_n \middle| X\right]\mathbb{1}_{A_n} \leq \frac{2}{\delta_n}\left[E_{f_0}KL(\Psi\|\Pi[\cdot\,|X]) + Ce^{-\delta_n/2}\right]$$

Idea: Apply duality formula

$$KL(Q\|P) = \sup_{\phi}\int \phi\, dQ - \log\int e^{\phi}\, dP$$

to $\phi(f) = \frac{1}{2}\delta_n\mathbb{1}_{\mathcal{F}_n}(f)$

Expected KL

[Nieman et al. '22] For suitable GP prior (with good RKHS approximations), the variational posterior satisfies

$$E_{f_0} KL(\Psi \| \Pi[\cdot\, |X]) \lesssim n r_n^* E_{f_0} \left\| K_{nn} - K_{nq} K_{qq}^{-1} K_{qn} \right\| + E_{f_0} tr \left(K_{nn} - K_{nq} K_{qq}^{-1} K_{qn} \right)$$

Also, [Shawe-Taylor & Williams, '02]

$$E_x \underbrace{\sum_{j=j_0}^{n} \mu_j}_{\text{spectrum of } K_{nn}} \leq n \underbrace{\sum_{j=j_0}^{\infty} \tilde{\lambda}_j}_{\text{spectrum of } \mathcal{A}^* \Lambda \mathcal{A}}$$

Inducing points

Theorem (Posterior contraction)

For $f_0 \in \bar{H}^\beta$ and

1. *Mildly:* $\lambda_i \asymp i^{-1-2\beta}$ *and* $q \geq n^{1/(1+2\beta+2p)}$, *for* β *large*
2. *Severely:* $\lambda_i \asymp i^{-\alpha} e^{-\xi i^p}$ *and* $q^p \geq (c + 2\xi)^{-1} \log n$

The variational posterior contracts at the minimax L^2*-rate* r_n^*.

Because of slow rates, small number of inducing variables needed (smaller for bigger degrees of ill-posedness)

Simulations

Heat equation

Recovery of the initial condition

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t), \quad u(x,0) = f_0(x), \quad u(0,t) = u(1,t) = 0$$

from observations of $\mathcal{A}f_0(x) = u(x,T)$.

- $\mathcal{A}\colon L^2[0,1] \mapsto L^2[0,1]$
- $\mathcal{A}f(x) = \sqrt{2}\sum_{i=1}^{\infty} f_i e^{-i^2\pi^2 t}\sin(i\pi x)$ for $f_i = \sqrt{2}\int_0^1 f(s)\sin(i\pi s)ds$

Heat equation

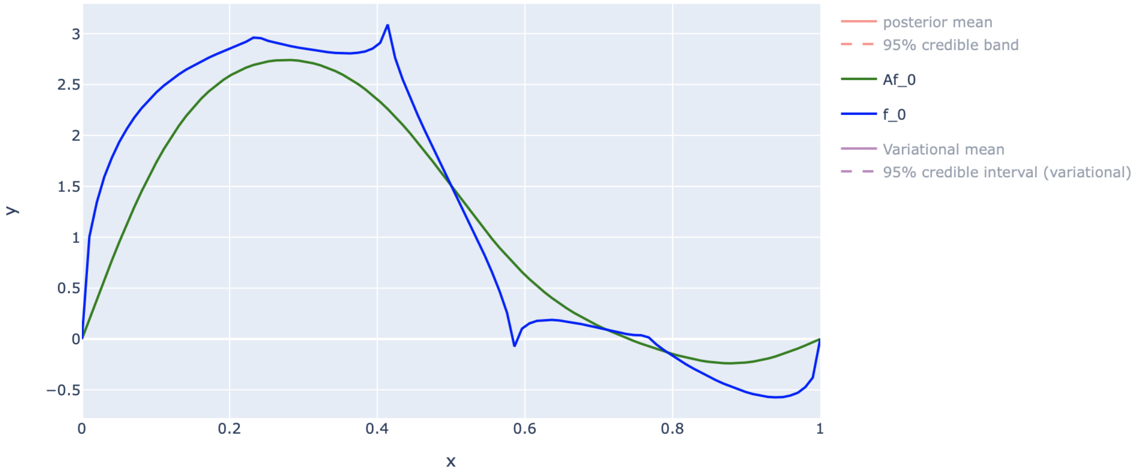
GP prior

$$W = \sqrt{2}\sum_{i=1}^{\infty} e^{-\xi i^2/2} Z_i \sin(i\pi x)$$

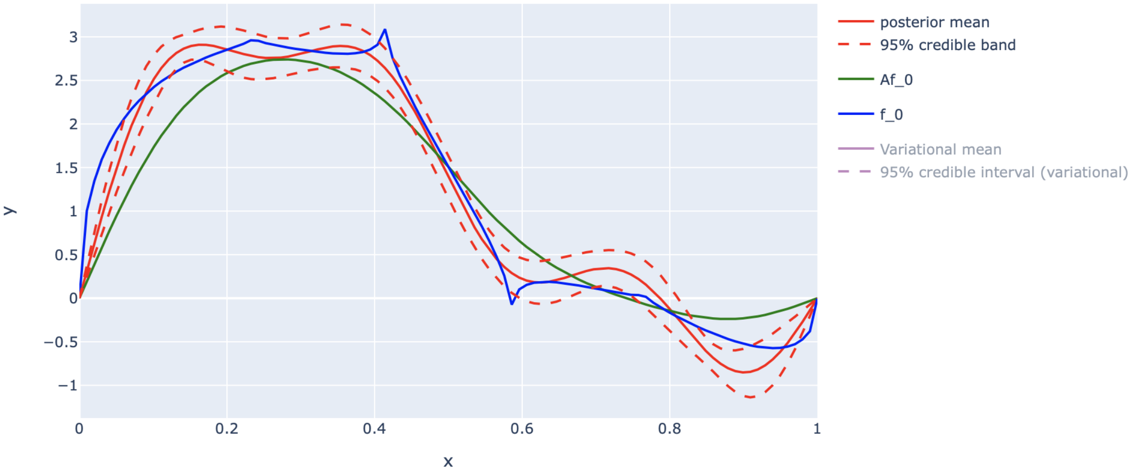
- $n = 8000$, $T = 5.10^{-3}$, $\xi = 10^{-1}$
- $q \asymp \left(\xi + 2 * (\pi^2) T)^{-1} \log(n)\right)^{1/2}\right) = 7$
- For f_0, we choose $\beta = 0.5$ and

$$f_{0,i} = \begin{cases} \left(1 + 0.4 * \sin(\sqrt{5}\pi i)\right) \; i^{-(\beta+1)} & \text{if} \quad i \text{ even} \\ \left(2.5 + 2 * \sin(\sqrt{2}\pi i)\right) \; i^{-(\beta+1)} & \text{if} \quad i \text{ odd} \end{cases}$$

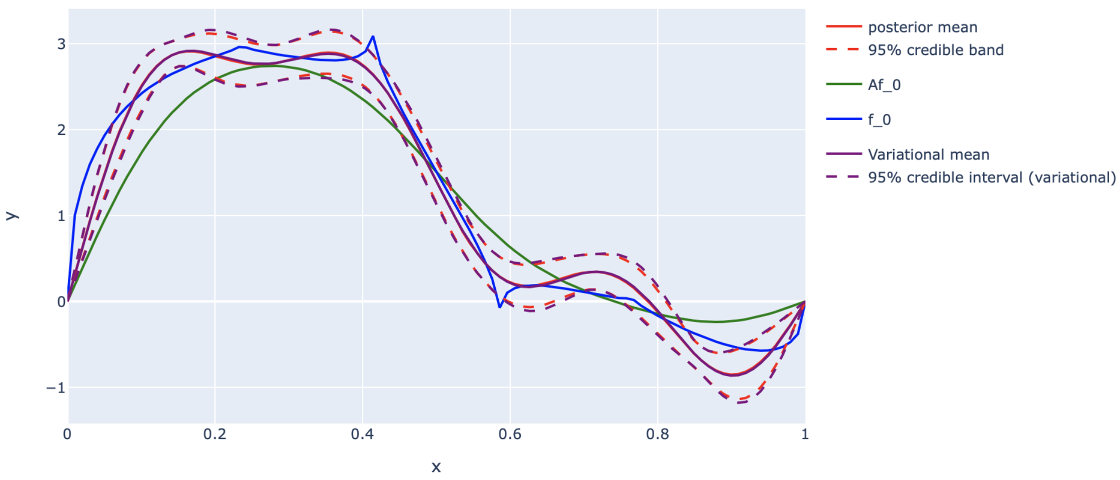
Heat equation



Heat equation



Heat equation



Conclusion

- We do not need vanishing KL for good variational posterior results in inverse problems
- Depending on degree of ill-posedness, need for logarithmic to sublinear number of inducing variables

- **Next:** What if eigenbases of $\mathcal{A}$ and Λ do no match ? Deconvolution ?

Thanks !