A Variational Inference Approach to Sparse Linear Models with Gamma Hyperpriors

Hwanwoo Kim

Joint work with S.Agrawal, D.Sanz-Alonso, A.Strang

University of Chicago

2022 ISBA Meeting

June 29, 2022

Problem Setup

Given data y of the form

$$y = Au + \eta, \quad \eta \sim \mathcal{N}(0, \Gamma),$$

where $A \in \mathbb{R}^{n \times d}$ and $\Gamma \in \mathbb{R}^{n \times n}$ are known,

- Estimation of u under sparsity assumption?
- Uncertainty quantification of u?

Outline

1. Hierarchical Bayesian Model
2. Iterative Alternating Scheme (IAS)
3. Variational Iterative Alternating Scheme (VIAS)
4. Illustrative Example

Hierarchical Bayesian Model

- Data generating model (Likelihood) $\implies y|u \sim \mathcal{N}(Au, \Gamma)$
- Bayesian framework $\implies u|\theta \sim \mathcal{N}(0, D_\theta),\ D_\theta = \mathrm{diag}(\theta)$
- Hierarchical setup $\implies \theta_i \sim \mathrm{Gamma}(\alpha_i, \beta),\ 1 \leq i \leq d$
- Ultimate objective: Posterior distribution $p(u, \theta|y)$

Posterior Distribution

$$p(u,\theta|y) = \frac{p(y|u,\theta)p(u|\theta)p(\theta)}{p(y)} \propto \exp\big(-J(u,\theta)\big),$$

where

$$J(u,\theta) := \overbrace{\frac{1}{2}\|y - Au\|_\Gamma^2 + \rlap{$\underbrace{\phantom{\frac{1}{2}\|u\|_{D_\theta}^2 + \sum_{i=1}^d \left[\frac{\theta_i}{\alpha_i} - \left(\beta - \frac{3}{2}\right)\log\frac{\theta_i}{\alpha_i}\right]}}_{(b)}$}\frac{1}{2}\|u\|_{D_\theta}^2}^{(a)} + \textstyle\sum_{i=1}^d \left[\frac{\theta_i}{\alpha_i} - \left(\beta - \frac{3}{2}\right)\log\frac{\theta_i}{\alpha_i}\right].$$

Iterative Alternating Scheme (IAS): Estimation of u

1. **Initialize** $\theta^0,\ k = 0.$
2. **Iterate until convergence:**
   1. Update

$$
\begin{aligned}
u^{k+1} &= \arg\min_u J(u, \theta^k) \\
&= \arg\min_u \frac{1}{2}\|y - Au\|_\Gamma^2 + \frac{1}{2}\|u\|_{D_\theta^k}^2 \\
&= (A^\top \Gamma^{-1} A + D_{\theta^k}^{-1})^{-1} A^\top \Gamma^{-1} y.
\end{aligned}
$$

   2. Update

$$
\begin{aligned}
\theta^{k+1} &= \arg\min_\theta J(u^{k+1}, \theta) \\
&= \alpha_i \left(\frac{\tilde{\beta}}{2} + \sqrt{\frac{\tilde{\beta}^2}{4} + \frac{(u_i^{k+1})^2}{2\alpha_i}} \right), \quad \tilde{\beta} = \beta - 3/2.
\end{aligned}
$$

   3. $k \to k + 1.$

More details in [Calvetti et al., 2019]

Laplace Approximation: Uncertainty Quantification of u

Approximate $p(u,\theta|y)$ by a Gaussian distribution with

- mean: $z_{lp} = \arg\min_{z=(u,\theta)} J(z)$
- covariance: $\nabla^2 J(z_{lp})^{-1}$ where

$$\nabla^2 J(z_{lp}) = \begin{bmatrix} A^\top \Gamma^{-1} A + \mathsf{diag}(1/\theta) & -\mathsf{diag}(u/\theta^2) \\ -\mathsf{diag}(u/\theta^2) & \mathsf{diag}(u^2/\theta^3 + \tilde{\beta}/\theta^2) \end{bmatrix}.$$

Mean-Field Variation Inference

- Can we approximate $p(u, \theta|y)$ with other probability distributions?
  $\implies$ Mean-Field Variation Inference [Blei et al., 2017]
- Mean-Field Family:

$$\mathcal{D} := \Big\{ q(z) : q(u,\theta) = q(u)q(\theta), \quad q(\theta) = \prod_{i=1}^{d} q(\theta_i) \Big\}.$$

- Find a distribution $q^* \in \mathcal{D}$ closest to $p(u, \theta|y)$.
  $\implies$ Kullback-Leibler Divergence

$$\begin{aligned} \mathsf{KL}(q(\cdot)||p(\cdot|y)) &:= \int q(z) \log \frac{q(z)}{p(z|y)} dz \\ &= \log p(y) - \underbrace{\int q(z) \log \frac{p(z,y)}{q(z)} dz}_{=:\mathsf{ELBO}(q)} \end{aligned}$$

Variational Iterative Alternating Scheme (VIAS)

1. **Initialize** $q^0(u)$, $k = 0$.
2. **Iterate:**
   1. (i) Update
   $$q^{k+1}(\theta) = \arg\max_{q(\theta)} \mathsf{ELBO}\big(q^k(u)q(\theta)\big)$$
   2. (ii) Update
   $$q^{k+1}(u) = \arg\max_{q(u)} \mathsf{ELBO}\big(q(u)q^{k+1}(\theta)\big)$$
   3. (iii) $k \to k+1$

VIAS: $q(\theta)$ update

With

$$q^{k+1}(\theta) = \prod_{i=1}^{d} q^{k+1}(\theta_i), \quad m_i^k = \mathbb{E}_{q^k(u)}[u_i], \quad C_{ii}^k = \mathsf{Var}_{q^k(u)}[u_i],$$

one can derive

$$q^{k+1}(\theta_i) = \mathsf{GIG}(b, r_i^k, s)$$

where $b = 2\beta$, $s = \alpha - 0.5$ and $r_i^k = (m_i^k)^2 + C_{ii}^k$.

VIAS: $q(u)$ update

One can obtain

$$q^{k+1}(u) = \mathcal{N}(m^{k+1}, C^{k+1})$$

with

$$\begin{aligned} m^{k+1} &= (A^\top \Gamma^{-1} A + L)^{-1} A^\top \Gamma^{-1} y, \\ C^{k+1} &= (A^\top \Gamma^{-1} A + L)^{-1}, \end{aligned}$$

where $L = \text{diag}\left(\mathbb{E}_{q^{k+1}(\theta)}\left[\frac{1}{\theta}\right]\right)$.

VIAS: Parameter Update Persepective

Input: Data y, matrix A. Prior hyperparameters: α, β.
Initialize: m^0, C^0. Set $b = 2\beta$, $s = \alpha - 0.5$.
Until convergence:

1. Update $r_i^{k+1} = (m_i^k)^2 + C_{ii}^k$ for each $i = 1, \ldots, d$.
2. Set

$$L = \mathrm{diag}(\ell), \qquad \ell_i = \frac{K_{s-1}\left(\sqrt{r_i^{k+1} b}\right)}{K_s\left(\sqrt{r_i^{k+1} b}\right)} \cdot \sqrt{\frac{b}{r_i^{k+1}}},$$

   and update

$$\begin{aligned} m^{k+1} &= (A^\top \Gamma^{-1} A + L)^{-1} A^\top \Gamma^{-1} y, \\ C^{k+1} &= (A^\top \Gamma^{-1} A + L)^{-1}. \end{aligned}$$

VIAS: Convergence and Initialization

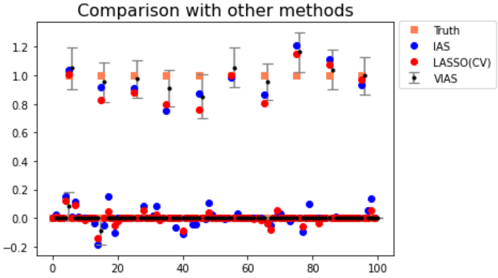
- ELBO(q) $=$ ELBO(m, C, r).
- ELBO(m, C, r) is not convex.
- In general, VIAS can only guarantee convergence to a local maximum.
- Due to multiple local maxima, choosing C^0 with small condition number is recommended.
- Hyperparameters α, β can be chosen to maximize ELBO
  : Only need ballpark values.
- More details in [Agrawal et al., 2022].

Comparison with other methods

Given

$$y = Au + \eta, \quad \eta \sim \mathcal{N}(0, \gamma^2 I_{50}),$$

where $A \in \mathbb{R}^{50\times 100}$, $\gamma = 0.02\|Au\|_\infty$.



Learning parameters for Lorenz 63 model I

Consider the Lorenz-63 system:

$$\frac{dx}{dt} = \sigma(y - x),$$
$$\frac{dy}{dt} = x(\rho - z) - y,$$
$$\frac{dz}{dt} = xy - \zeta z,$$

with the classical parameter values $\sigma = 10, \rho = 28$, $\zeta = 8/3$.

Learning parameters for Lorenz 63 model II

Given data $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$, a representation of each trajectory for time interval [0, 40] along 2000 equidistant time points, construct a matrix of the form:

$$A = \begin{bmatrix} | & | & | & | & | & | & | & | & | & & | & \\ \mathbf{x} & \mathbf{y} & \mathbf{z} & \mathbf{x}^2 & \mathbf{y}^2 & \mathbf{z}^2 & \mathbf{xy} & \mathbf{xz} & \mathbf{yz} & \dots & \mathbf{x}^5 & \dots \\ | & | & | & | & | & | & | & | & | & & | & \end{bmatrix} \in \mathbb{R}^{2000\times 55}.$$

Learning parameters for Lorenz 63 model III

Obtain synthetic data $\dot{\mathbf{x}}$, $\dot{\mathbf{y}}$, $\dot{\mathbf{z}}$ from

$$\begin{aligned} \dot{\mathbf{x}} &= A\Phi_1 + \eta_1, \quad \Phi_1 = [-10, 10, 0, 0, 0, 0, 0, \dots, 0]^\top, \\ \dot{\mathbf{y}} &= A\Phi_2 + \eta_2, \quad \Phi_2 = [28, -1, 0, 0, -1, 0, 0, \dots, 0]^\top, \\ \dot{\mathbf{z}} &= A\Phi_3 + \eta_3, \quad \Phi_3 = [0, 0, -8/3, 0, 0, 0, 1, \dots, 0]^\top, \end{aligned}$$

where $\eta_i \sim \mathcal{N}(0, 0.3 I_{2000})$ are independent.

- Goal: recover Φ_1, Φ_2, Φ_3 based on A, $\dot{\mathbf{x}}$, $\dot{\mathbf{y}}$, $\dot{\mathbf{z}}$.
- More detailed explanation on simulation setup: [Brunton et al., 2016]

Parameter Estimation and Uncertainty Quantification

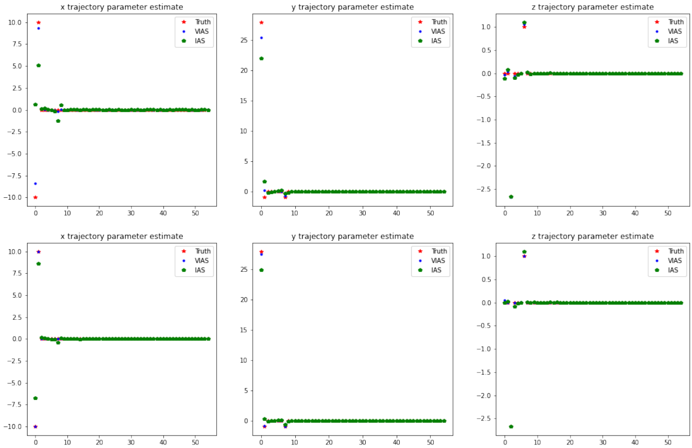


Figure: Recovery of dictionary coefficients for x-trajectory (first column), y-trajectory (second column), and z-trajectory (third column) using IAS and VIAS. Top: two iterations. Bottom: five iterations.

Uncertainty Quantification of Trajectory: IAS

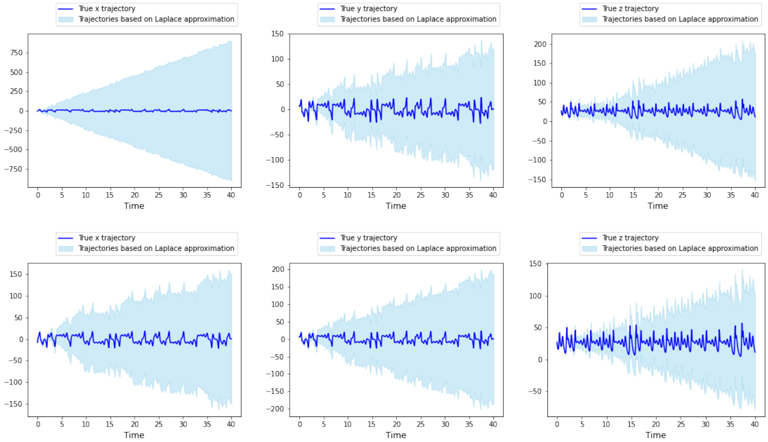


Figure: True Lorenz-63 trajectory and IAS estimation. Top: two IAS iterations. Bottom: five IAS iterations. Blue line is the true dynamics. Shaded regions are constructed from 2.5 and 97.5 credible levels of coefficients.

Uncertainty Quantification of Trajectory: VIAS

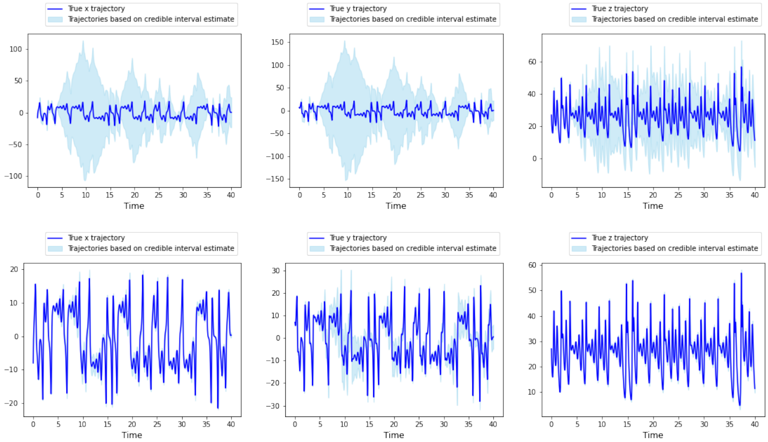


Figure: True Lorenz-63 trajectory and VIAS estimation. Top: two VIAS iterations. Bottom: five VIAS iterations. Blue line is the true dynamics. Shaded regions are constructed from 2.5 and 97.5 credible levels of coefficients.

Summary

- Hierarchical Bayesian model with gamma hyperprior to induce sparse structure in parameter of interest.
- Employed a variational inference technique to approximate posterior distribution of the target parameter.
- An coordinate ascent variational inference led to simple parameter update rules.
- Can generalize to nonlinear sparse regression setting with the same Hierarchical Bayesian model, but with ensemble-based derivative-free nonlinear optimization methods.

References

Agrawal, S., Kim, H., Sanz-Alonso, D., and Strang, A. (2022).
A variational inference approach to inverse problems with gamma hyperpriors.
arXiv preprint arXiv:2111.13329,
to appear in SIAM/ASA Journal of Uncertainty Quantification.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017).
Variational inference: A review for statisticians.
Journal of the American statistical Association, 112(518):859–877.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016).
Discovering governing equations from data by sparse identification of nonlinear dynamical systems.
Proceedings of the national academy of sciences, 113(15):3932–3937.

Calvetti, D., Somersalo, E., and Strang, A. (2019).
Hierachical bayesian models and sparsity: ℓ_2-magic.
Inverse Problems, 35(3):035003.

Kim, H., Sanz-Alonso, D., and Strang, A. (2022).
Hierarchical ensemble kalman methods with sparsity-promoting generalized gamma hyperpriors.
arXiv preprint arXiv:2205.09322.