The Infemus Estimator

Timothée Stumpf Fétizon [tim.stumpf-fetizon@warwick.ac.uk]

May 2, 2022

Joint work with O. Papaspiliopoulos (Bocconi), J. Weare (NYU)

Outline

Motivation

+

Bayesian Marginal Likelihood

We work with a *parametric statistical model* $p(y, \theta|\lambda)$, where:

- y is data
- θ are parameters (possibly high-d)
- λ are low-d hyperparameters

For the purpose of *model selection*, the normalizing constant of the posterior, often called the *marginal likelihood*

$$z(\lambda) = \int p(y, \theta|\lambda) \, \mathrm{d}\theta \tag{1}$$

is of particular interest, and usually intractable!

Maximization vs. Exploration

Many methods use marginal likelihood to set λ, solving

$$\underset{\lambda}{\operatorname{argmax}}\, z(\lambda) \tag{2}$$

Instead, we wish to fully characterize and explore

$$\lambda \mapsto z(\lambda) \tag{3}$$

on compact set Λ, e.g. to investigate bi-modality.

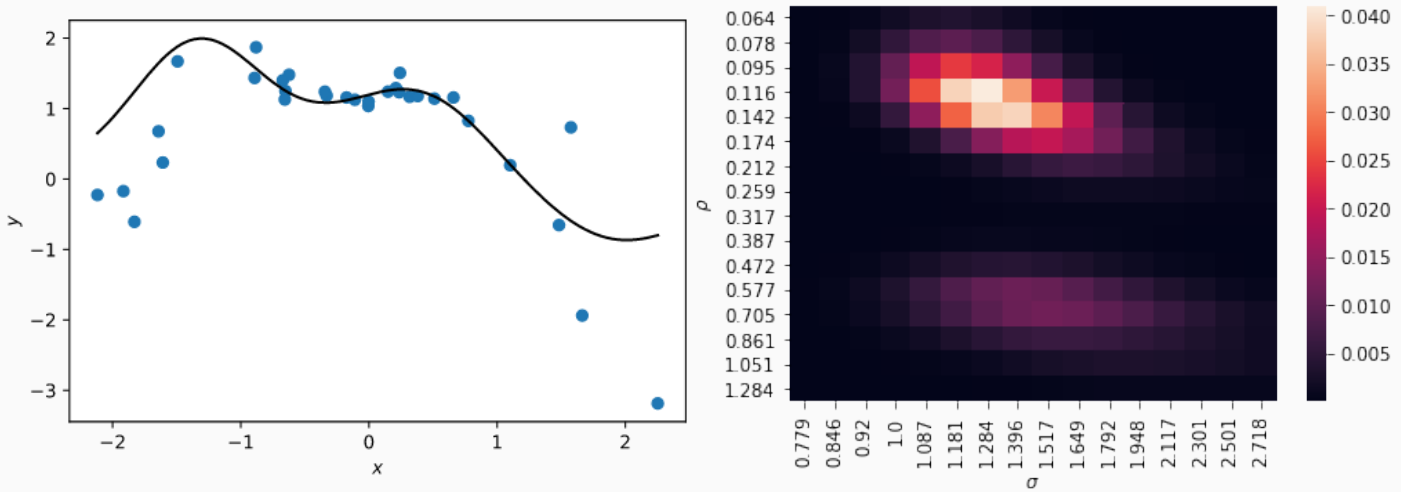

Figure 1: Data and marginal likelihood for hyperparameters of a GP regression.

Background: The EMUS Estimator

Importance Sampling

Consider identity

$$z(\lambda) = \int \underbrace{\frac{p(y, \theta|\lambda)}{q(\theta)}}_{\text{integrand}} \underbrace{q(\theta)}_{\text{sampling dens.}} \, \mathrm{d}\theta \tag{4}$$

and implied estimator

$$\hat{z}(\lambda) = n^{-1} \sum_{m=1}^{n} \frac{p(y, \theta^{(m)}|\lambda)}{q(\theta^{(m)})}, \qquad \theta^{(m)} \sim q(\theta) \tag{5}$$

Fix lattice of *windows* $(\lambda_1, \lambda_2, \dots, \lambda_\ell)$ and define mixture density

$$q(\lambda) = \sum_{i=1}^{\ell} z_i p(\theta_i|y, \theta) \tag{6}$$

for mixture weights z_i.

Linear System

This yields linear system

$$z_j = \sum_{i=1}^{\ell} z_i \underbrace{\int \frac{p(y, \theta|\lambda_j)}{\sum_{k=1}^{\ell} p(y, \theta|\lambda_k)} p(\theta|y, \lambda_i) \, \mathrm{d}\theta}_{\stackrel{\text{def}}{=} f_{ij} \leq 1} = \sum_{i=1}^{\ell} z_i f_{ij} \tag{7}$$

In linear algebra notation,

$$\mathbf{z} = \mathbf{F}^{\mathsf{T}} \mathbf{z} \tag{8}$$

where $\mathbf{F}$ is row stochastic, and $\mathbf{z}$ is a right eigenvector for $\mathbf{F}^{\mathsf{T}}$.

Side Note

Subject to $\mathbf{z}^{\mathsf{T}} \mathbf{1} = 1$, $\mathbf{z}$ is invariant for Markov chain with transition matrix $\mathbf{F}$!

EMUS Estimator

Given n_i samples $\theta_i^{(m)} \sim p(\theta|y, \lambda_i)$ from each posterior, we have

$$\hat{z}_j = \sum_{i=1}^{\ell} \hat{z}_i \underbrace{n_k^{-1} \sum_{m=1}^{n_k} \frac{p(y, \theta_i^{(m)}|\lambda_j)}{\sum_{k=1}^{\ell} p(y, \theta_i^{(m)}|\lambda_k)}}_{\stackrel{\text{def}}{=} \hat{f}_{ij} \leq 1} = \sum_{i=1}^{\ell} \hat{z}_i \hat{f}_{ij} \tag{9}$$

Integrand is bounded for all (i, j) (nice)! Estimator $\hat{\mathbf{z}}$ solves the linear system

$$\hat{\mathbf{z}} = \hat{\mathbf{F}}^{\mathsf{T}} \hat{\mathbf{z}} \tag{10}$$

Solve with constraint $\hat{\mathbf{z}}^{\mathsf{T}} \mathbf{1} = 1$. Weare et al. call this estimator *EMUS* (Eigenvector method for Umbrella sampling).

Analysis of EMUS

Estimator is well understood:

- Similar equations give rise to fixed-point iteration which converges to *Vardi's estimator* [Weare et al., 2020]. It is asymptotically efficient [Tan, 2004]...
- But EMUS is easier to analyze! For $\mathbf{H} = \mathbf{I} - \mathbf{F}$,

$$\mathrm{Cov}\left[\hat{\mathbf{z}} - \mathbf{z}\right] \approx \mathbf{H}^{\mathsf{T}\#} \left(\sum_{i=1}^{\ell} z_i^2 \, \mathrm{Cov}\left[\hat{\mathbf{f}}_i\right] \right) \mathbf{H}^{\#}, \qquad \mathbf{H} = \mathbf{I} - \mathbf{F} \tag{11}$$

  where $\mathbf{H}^{\#}$ is the *group inverse*. This is asymptotically exact.

Some interesting angles...

- **Experimental Design**: The n_is are tuning parameters and can be optimized to reduce asymptotic variance.
- **Perturbation Bounds**: Theory of Markov chains yields intuitive bounds on asymptotic variance.

Novelty: Inf-dim EMUS

EMUS Off The Lattice

EMUS is well-behaved, enjoys CLT. But when estimating surfaces, additional considerations apply...

- **Interpolation**: We'd like to fill in the function off the lattice.
- **Smoothness**: $\lambda \mapsto z(\lambda)$ is typically smooth, which we haven't used at all.

We address those with Bayesian nonparametrics. Assume that a priori

$$z \sim \mathrm{GP}\left[\eta, \kappa\right] \tag{12}$$

where κ imposes smoothness and a fixed-sum constraint. We consider 2 ways of combining EMUS with the prior...

Smoothing the Operator

We postulate the soft constraint

$$\hat{\mathbf{H}}^\top \mathbf{z}_{\text{grid}} \approx \hat{\mathbf{H}}^\top \hat{\mathbf{z}}, \qquad \mathbf{z}_{\text{grid}} = (z(\lambda_i))_{i \in \text{grid}} \tag{13}$$

Implement this with the observational model

$$\hat{\mathbf{H}}^\top \hat{\mathbf{z}} | z \sim \mathrm{N}\left[\hat{\mathbf{H}}^\top \mathbf{z}_{\text{grid}}, \mathbf{R}\right] \tag{14}$$

with $\mathbf{R}$ encoding fixed-sum constraint, and "observe" $\hat{\mathbf{H}}^\top \hat{\mathbf{z}} = \mathbf{0}$. z still is a Gaussian process a posteriori!

Side Note

Some numerical complications arise since $\mathbf{R}$ is only PSD...

Smoothing the Solution

So what's a good choice of $\mathbf{R}$? Recall asymptotic covariance of EMUS and set

$$\mathbf{R} = \sum_{i \in \text{grid}} z^2(\lambda_i) \operatorname{Cov}\left[\hat{\mathbf{f}}_i\right] \tag{15}$$

To keep GP tractable, need some plug-in estimator for $\mathbf{z}_{\text{grid}}$. This will come back to haunt us... But then, what about instead using EMUS as data:

$$\hat{\mathbf{z}}|z \sim \mathrm{N}\left[\mathbf{z}_{\text{grid}}, \hat{\mathbf{H}}^{\mathsf{T}\#}\mathbf{R}\hat{\mathbf{H}}^{\#}\right] \tag{16}$$

Again this implies that z is a GP a posteriori!

Side Note

Operator smoothing formulation doesn't need group inverse $\mathbf{H}^{\#}$, which could be numerically beneficial.

Some Output in a Hard Setting

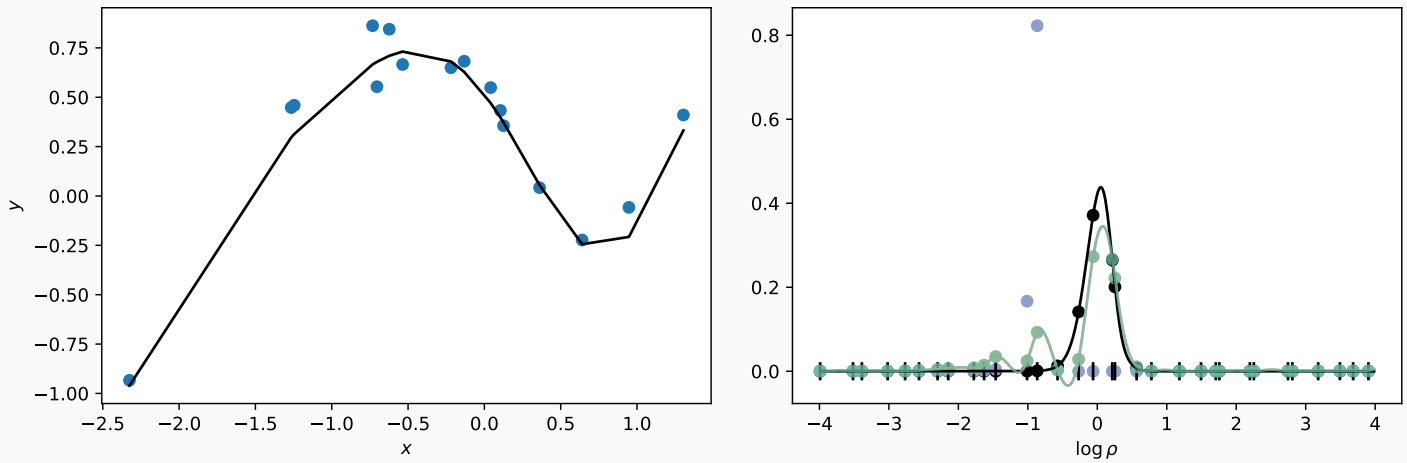

Figure 2: (a) Data of a GP regression. (b) In black, marginal likelihood for length scale of the GP kernel. In blue, EMUS estimate. In green, soft constraint estimator. In red, EMUS smoothing estimator.

So Are They The Same?

Proposition: Suppose $\mathbf{R}$ has 0-sum rows and columns, i.e. it is the covariance matrix of a fixed-sum vector. Then both options imply the same posterior distribution on z!

Proof idea: Show observational models and corresponding observations are equivalent. Use $\mathbf{R}\hat{\mathbf{H}}^{\#}\hat{\mathbf{H}} = \mathbf{R}$.

Notice

This does *not* apply when $\mathbf{R}$ has full rank.

Infemus in Practice

Plug-In Covariance Estimate

Since

$$\mathbf{R} = \sum_{i \in \text{grid}} z^2(\lambda_i) \operatorname{Cov}\left[\hat{\mathbf{f}}_i\right] \tag{17}$$

depends on unknown function z, GP is no longer tractable. We carry out experiments with plug-in values

- Flat $\mathbf{z}_{\text{grid}} \propto \mathbf{1}$
- Oracle $\mathbf{z}_{\text{grid}}$.

Further operational details:

- Prior mean is flat.
- Kernel is Matern 5/2 with a fixed-sum constraint.
- Hyperparameters are set by an empirical Bayes procedure.

Don't Get Confused

There are two GPs: The data model, and the Infemus estimator.

Some Output

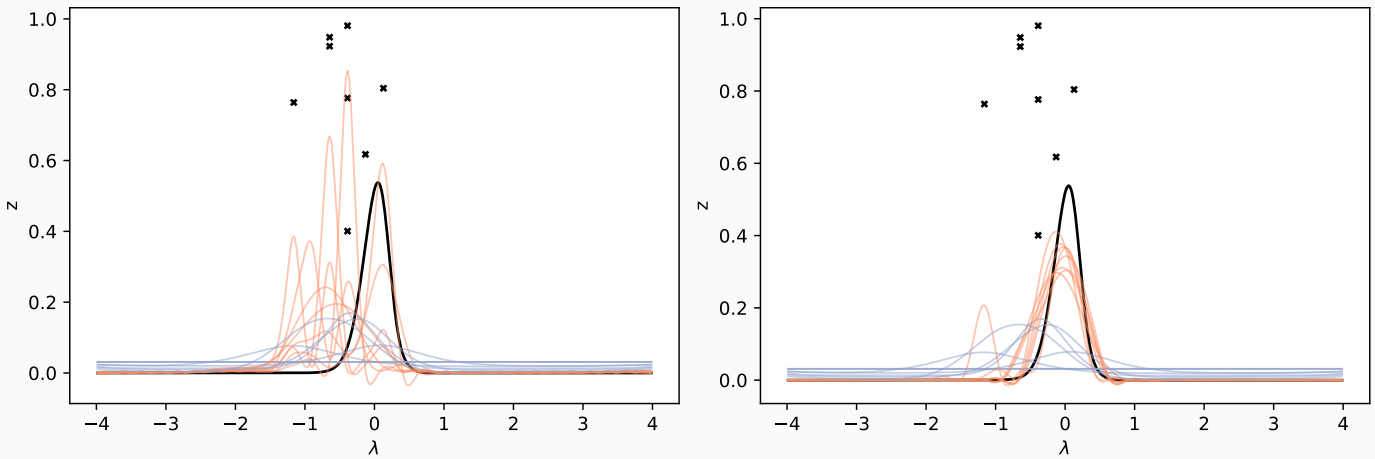

Figure 3: Some Monte Carlo samples of Infemus (red), isotropic GP smoothing (blue), and corresponding EMUS modes as crosses. (a) Flat plug-in. (b) Oracle plug-in.

Iterative Scheme

Should we try to infer the correct non-Gaussian posterior?

- (+) Once $\mathbf{z}_{\text{grid}}$ is known, z is a GP.
- (+) Solution provides an upper pound on what's feasible.
- (–) Requires inference through MCMC. Substituting a hard computational problem for another invites absurdity...

Is there a simple approximating scheme? Try iteration

1. Set $\mathbf{R}^{(0)}$ using an initial guess for $\mathbf{z}_{\text{grid}}^{(0)}$.
2. Find posterior mean $\mathbf{z}_{\text{grid}}^{(1)}$. Set $\mathbf{R}^{(1)}$ using $\mathbf{z}_{\text{grid}}^{(1)}$.
3. And so forth. Maybe there's a good fixed point?

Some Output

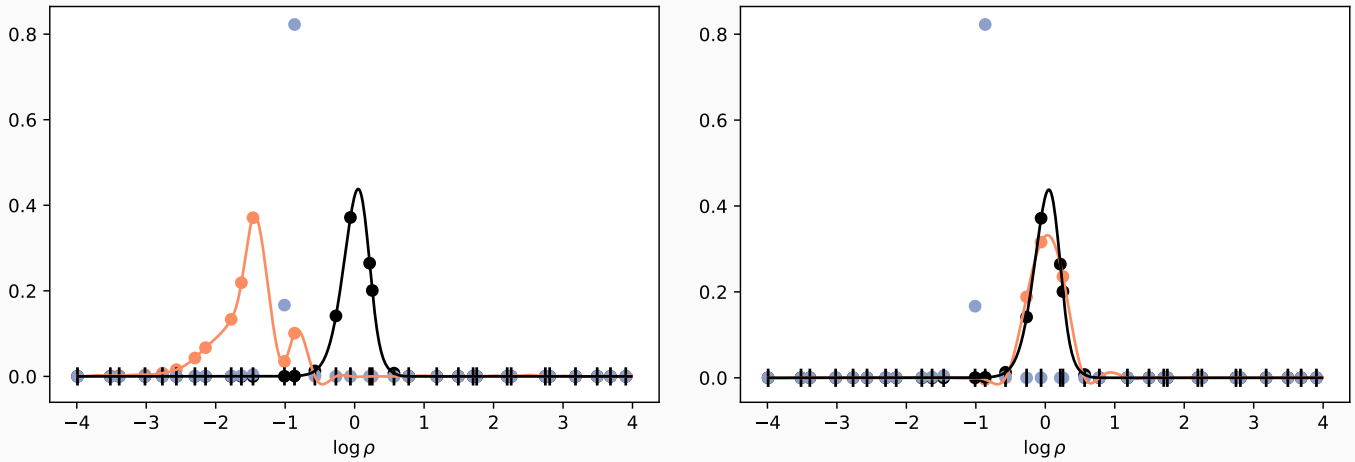

Figure 4: Estimate after 100 iterations. (a) Flat plug-in at iteration 0. (b) Oracle plug-in at iteration 0.

WIP...

- *How* hard is inference for the correct model? Preliminary experiments suggests the objective is large around ground truth...
- Is there a reliable iterative scheme which optimizes sampling intensity for each row of $\mathbf{F}$?
- The other side of the model selection coin: Can we use EMUS to produce good estimators of *Bayesian predictive criteria*?

Stay tuned for the pre-print!