Gaussian process optimisation for approximate Bayesian inference

Submitted to Journal of Econometrics

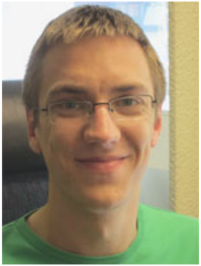
Johan Dahlin
johan.dahlin@liu.se
work.johandahlin.com

Division of Automatic Control,
Linköping University, Sweden.

LINKÖPING UNIVERSITY

This is collaborative work together with

Prof. Mattias Villani (Linköping University, Sweden)
Dr. Fredrik Lindsten (University of Cambridge, United Kingdom).
Prof. Thomas Schön (Uppsala University, Sweden).



Why are we doing this?

- Bayesian parameter inference for general SSMs.
- Accelerate existing methods for estimating $\pi(\theta)$.
- Simplify tuning of posterior sampling algorithms.

How will we do this?

- Make use of prior knowledge of $\pi(\theta)$.
- Construct a surrogate of $\pi(\theta)$ using noisy samples.
- Extract a Laplace approximation around its mode. (BvM)

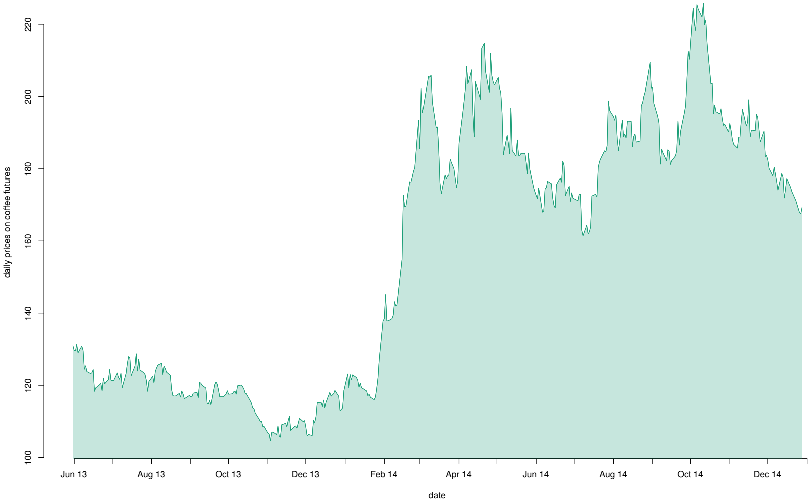
What are we going to do?

- Introduce models with intractable likelihoods.
- Discuss Gaussian processes and particle filtering.
- Benchmark with alternatives (sampling methods).

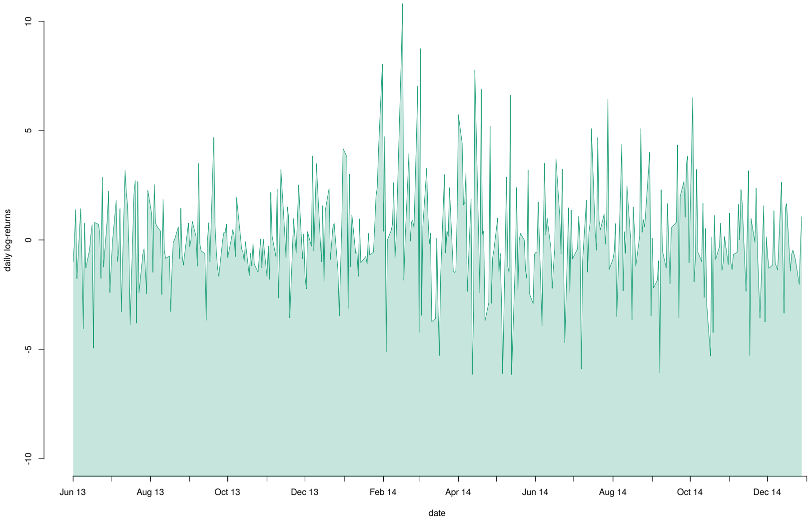
Motivating example: coffee futures

Credits: wordridden@FlickR

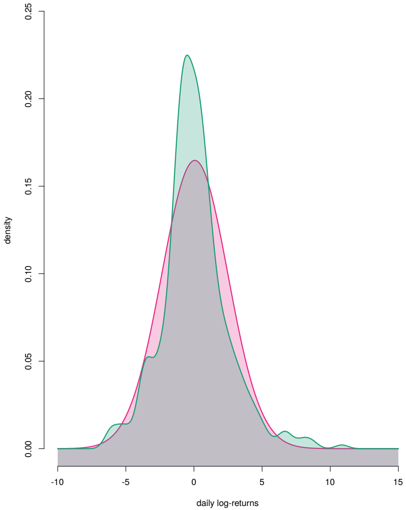
Motivating example: coffee futures

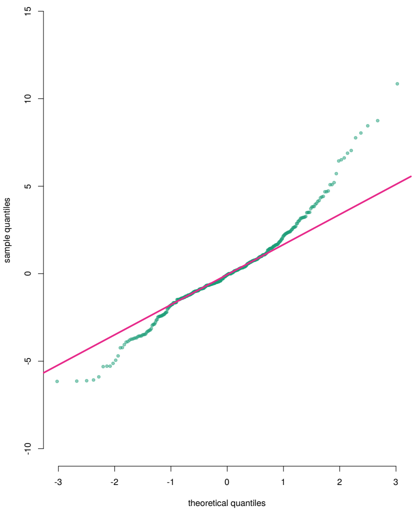


Motivating example: coffee futures



Motivating example: coffee futures





An SSM with intractable likelihoods

Consider the stochastic volatility model given by:

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \mu + \phi(x_t - \mu), \sigma_v\Big),$$
$$y_t|x_t \sim \mathcal{A}\Big(y_t; \alpha, \exp(x_t)\Big).$$

with parameters $\theta = \{\mu, \phi, \sigma_v, \alpha\}$. The use of $\mathcal{A}(\alpha, \sigma)$ is motivated by:

- Possibly heavy tailed and skewed.
- Family of stable distributions (closed under affine transformations).
- Generalised central limit theorem (infinite second moments).
- Lévy-Itō decomposition (discrete time analogue to a Lévy process).

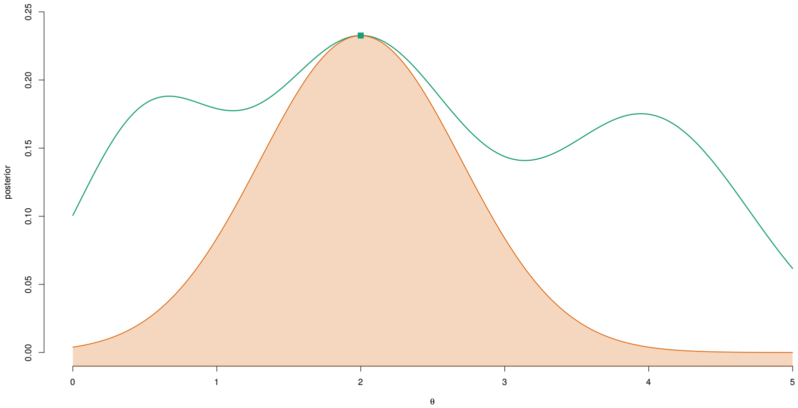
Parameter inference

- The likelihood $p_\theta(y_{1:T})$ and the posterior $\pi(\theta)$ are intractable.
- States $x_{0:T}$ are unknown and the pdf $\mathcal{A}(\alpha, \sigma)$ does not exist.
- In principle, we have

$$\pi(\theta) = p(\theta|y_{1:T}) = \frac{p_\theta(y_{1:T})p(\theta)}{p(y_{1:T})},$$

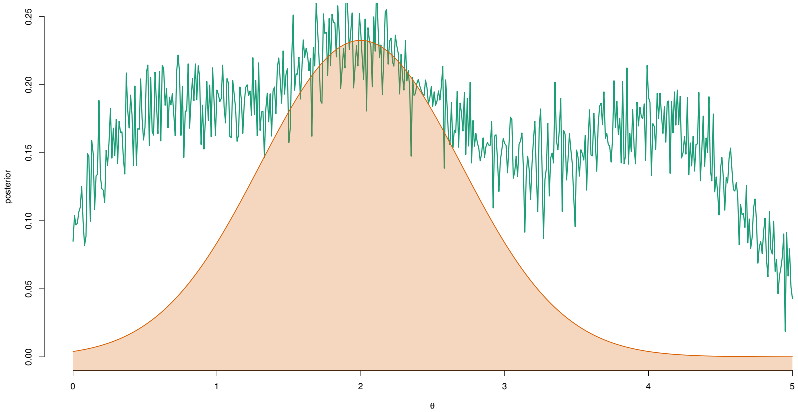
  but we need to approximate it. How? Sampling or analytical.
- Gaussian (Laplace) approximation of $\pi(\theta)$. BvM theorem.

Approximate parameter inference



$$\widehat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \log \pi(\theta). \quad \widehat{\pi}_{\text{Laplace}}(\theta) = \mathcal{N}\left(\widehat{\theta}_{\text{MAP}}, \left[-\nabla^2 \log \pi(\theta)\Big|_{\theta=\widehat{\theta}_{\text{MAP}}} \right]^{-1}\right).$$

Approximate parameter inference



$$\widehat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \log \pi(\theta). \quad \widehat{\pi}_{\text{Laplace}}(\theta) = \mathcal{N}\left(\widehat{\theta}_{\text{MAP}}, \left[-\nabla^2 \log \pi(\theta)\Big|_{\theta=\widehat{\theta}_{\text{MAP}}} \right]^{-1}\right).$$

Gaussian process optimisation

- An instance of Bayesian optimisation.
- Global optimisation method.
- Few function evaluations required and no gradients.
- Sampling/computing the objective function may take hours/days.

References:

J. Mockus, V. Tiesis and A. Zilinskas, The application of Bayesian methods for seeking the extremum. In L.C.W. Dixon and G.P. Szegos (editors), Toward Global optimisation, North-Holland, 1978.

D.J. Lizotte, Practical Bayesian optimisation. PhD thesis, University of Alberta, 2008.

J. Snoek, H. Larochelle and R.P. Adams, Practical Bayesian optimisation of Machine learning algorithms. In Advances in Neural Information Processing Systems (NIPS) 25, Curran Associates Inc, 2012.

Overview of Gaussian process optimisation

(i) Given iterate θ_k, estimate the log-posterior $\widehat{\pi}_k \approx \pi(\theta_k)$.

Particle filtering with approximate Bayesian computations.

(ii) Given $\{\theta_j, \widehat{\pi}_j\}_{j=0}^{k}$, create a surrogate function of $\pi(\theta)$.

Gaussian process predictive distribution.

(iii) Select a new θ_{k+1} using the surrogate function.

Acquisition function based on the Gaussian process posterior.

Particle filtering

- Can estimate $p_\theta(y_{1:T})$ with Gaussian errors (variance $\propto 1/N$).
- Approximate Bayesian computations. We cannot evaluate

$$w_t^{(i)} = g_\theta\Big(y_t | x_t^{(i)}\Big) = \mathcal{A}\Big(y_t; \alpha, \exp\left(x^{(i)}\right)\Big),$$

  but instead we use the approximation

$$y_t^\star \sim \mathcal{A}(\alpha, \exp(x_t)),$$
$$w_t^{(i)} = \mathcal{N}\Big(y_t; y_t^{(i),\star}, \epsilon^2\Big), \qquad \epsilon > 0.$$

- Computational time is usually several minutes.

Gaussian process regression [I/III]

- Estimator for $\pi(\theta)$ available with Gaussian error.

- A Gaussian process is an infinite dimensional Gaussian distribution.
- Prior specified by mean and covariance functions.
- Can be used for regression using a standard prior-posterior update.

Gaussian process regression [II/III]

We assume a priori

$$\pi(\theta) \sim \mathcal{GP}\Big(m(\theta), \kappa(\theta, \theta')\Big),$$

which together with the data

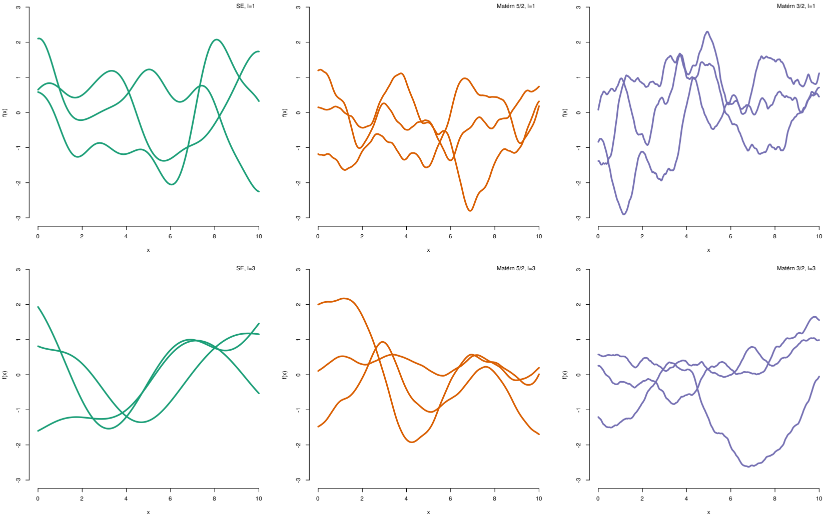
$$\widehat{\pi}(\theta) = \pi(\theta) + \sigma_{\widehat{\pi}} z_t, \qquad z_t \sim \mathcal{N}(0, 1),$$

gives the posterior predictive distribution

$$\pi(\theta_\star)|\mathcal{D}_k \sim \mathcal{N}\Big(\mu(\theta_\star|\mathcal{D}_k), \sigma^2(\theta_\star|\mathcal{D}_k) + \sigma^2_{\widehat{\pi}}\Big),$$

where the current data is denoted $\mathcal{D}_k = \{\theta_j, \widehat{\pi}_i\}_{j=1}^k$.

Gaussian process regression [III/III]



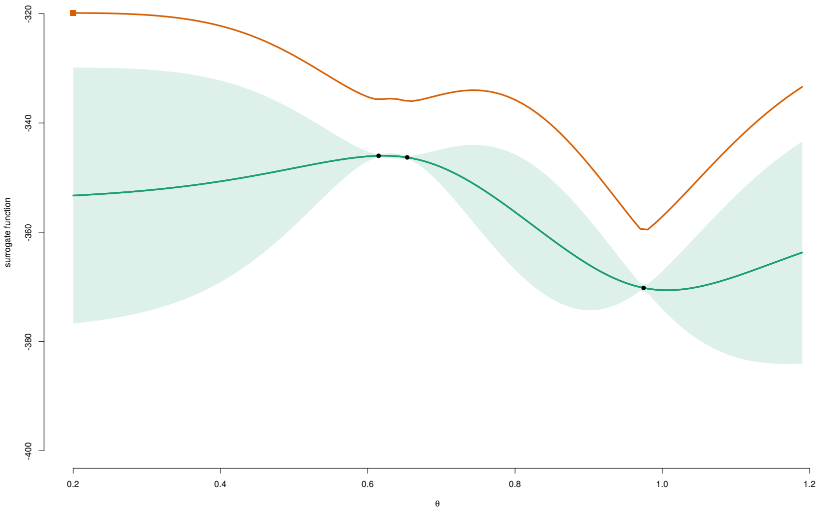
Acquisition rule for selecting sampling points

- Estimator for $\pi(\theta)$ available with Gaussian error.
- Surrogate of the log-posterior available as a Gaussian process.

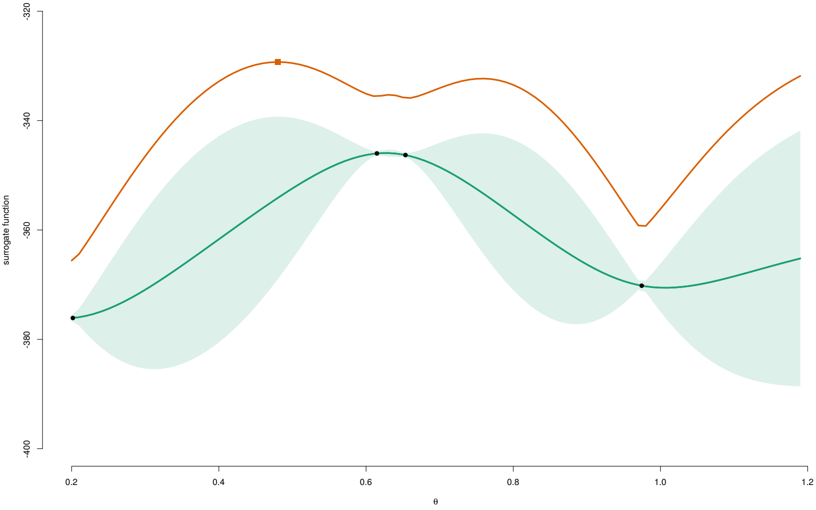
- Idea: use the Gaussian process model to select θ_{k+1}.
- Balance exploration (decrease the uncertainty) and exploitation (make use of the mean).
- We use the upper 95% confidence bound (UCB):

$$\theta_{k+1} = \underset{\theta_\star \in \Theta}{\operatorname{argmax}} \left[\mu(\theta_\star | \mathcal{D}_k) + 1.96 \sqrt{\sigma^2(\theta_\star | \mathcal{D}_k)} \right],$$

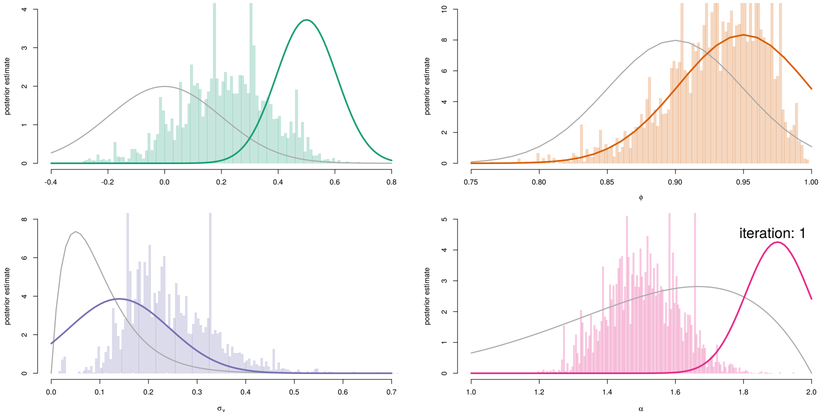
Example: Gaussian process optimisation [I/II]



Example: Gaussian process optimisation [II/II]



Motivating example: volatility of coffee futures [I/III]

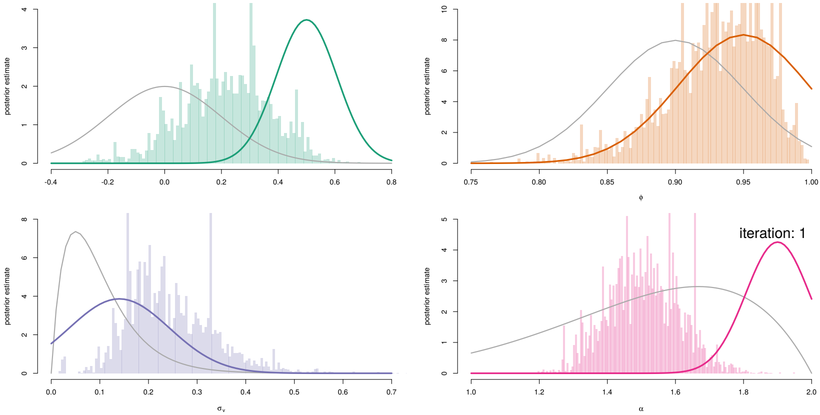


Stochastic volatility model with α-stable returns

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \mu + \phi(x_t - \mu), \sigma_v\Big), \qquad y_t|x_t \sim \mathcal{A}\Big(y_t; \alpha, \exp(x_t)\Big).$$

Log-posterior estimates for PMH-ABC: $30\,000$ (histogram) and GPO-ABC: 350 (solid lines).

Motivating example: volatility of coffee futures [II/III]

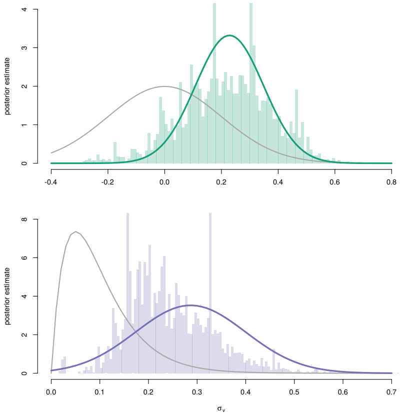


Stochastic volatility model with α-stable returns

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \mu + \phi(x_t - \mu), \sigma_v\Big), \qquad y_t|x_t \sim \mathcal{A}\Big(y_t; \alpha, \exp(x_t)\Big).$$

Log-posterior estimates for PMH-ABC: $30\,000$ (histogram) and GPO-ABC: 350 (solid lines).

Motivating example: volatility of coffee futures [III/III]

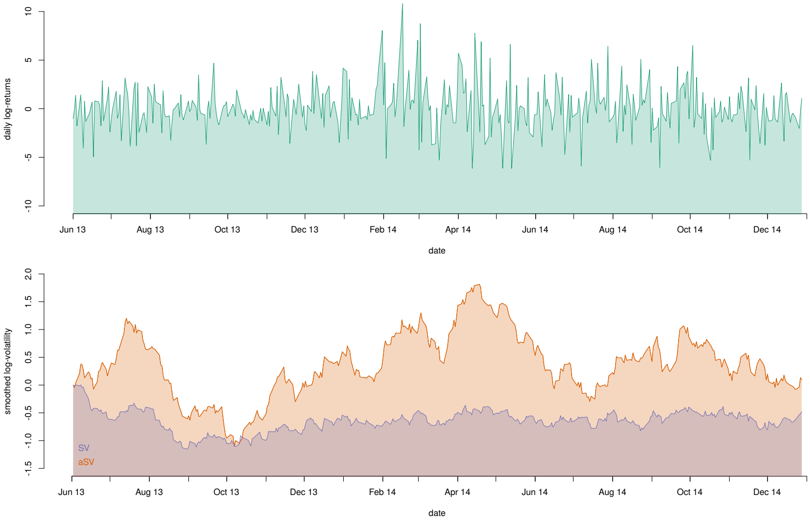


Stochastic volatility model with α-stable returns

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \mu + \phi(x_t - \mu), \sigma_v\Big), \qquad y_t|x_t \sim \mathcal{A}\Big(y_t; \alpha, \exp(x_t)\Big).$$

Log-posterior estimates for PMH-ABC: $30\,000$ (histogram) and GPO-ABC: 350 (solid lines).

Motivating example: volatility of coffee futures



Why are we doing this?

Bayesian parameter inference for general SSMs.

Accelerate existing methods for estimating $\pi(\theta)$.

Simplify tuning of posterior sampling algorithms.

How did we do this?

Gaussian process optimisation.

Particle filtering with approximate Bayesian computations.

Laplace approximations.

What did we achieve?

100 fold speed-up compared with PMH.

Speed-up compared with other optimisation methods.

Enables inference in copula models (Monday micro).

Estimate of the Hessian of $\log \pi(\theta)$ and its mode.

Thank you for listening

Comments, suggestions and/or questions?

Johan Dahlin

johan.dahlin@liu.se

work.johandahlin.com

LINKÖPING UNIVERSITY

References

J. Dahlin, M. Villani and T. B. Schön, Efficient approximate Bayesian inference for models with intractable likelihoods. Pre-print, arXiv:1506.06975v1, June 2015.

J. Dahlin and F. Lindsten, Particle filter-based Gaussian process optimisation for parameter inference. Proceedings of the 19th World Congress of the International Federation of Automatic Control (IFAC), Cape Town, South Africa, August 2014.

J. Dahlin, F. Lindsten and T. B. Schön, Quasi-Newton particle Metropolis-Hastings. Proceedings of the 17th IFAC Symposium on System Identification, Beijing, China, October 2015.