An ensemble filter for heavy-tailed t–distributions

Mathieu Le Provost[1], Ricardo Baptista[2],
Youssef Marzouk[3] and Jeff D. Eldredge[4]

[1]Long Island University, [2]California Institute of Technology,
[3]Massachusetts Institute of Technology, [4]University of California, Los Angeles

18th EnKF workshop
May 4, 2023

Motivations

Departure from Gaussian tails is a common feature of geophysical inference problems due to the nonlinear dynamical and observation processes and the uncertainty from the physical sensors.

Many filters like the EnKF assume **at least** that the **tails of the forecast distribution are Gaussian** and not suited for heavy-tailed distributions.

Objective: How can do consistent inference in heavy-tailed filtering problems?

Problem setting

We consider a generic state-space model:

The evolution of the state $(\mathbf{X}_t)_{t \geq 0}$ is fully described by the initial distribution $\pi_{\mathbf{X}_0}$ and the dynamical model:

$$\mathbf{X}_t = f(\mathbf{X}_{t-1}) + \mathbf{W}_t$$

We collect observations $(\mathbf{Y}_t)_{t>0}$ at every time step according to the observation model:

$$\mathbf{Y}_t = h(\mathbf{X}_t) + \mathcal{E}_t$$

Objective: Sequentially estimate **the filtering density** $\pi_{t|t} := \pi_{\mathbf{X}_t | \mathbf{Y}_{1:t} = y_{1:t}}$

Generic ensemble filtering algorithm

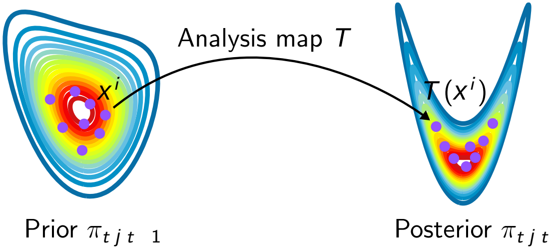
Ensemble filtering methods propagate a set of M particles $\{x^{(1)}, \dots, x^{(M)}\}$ to form an empirical approximation for the filtering density $\pi_{t\,|\,t}$.

1. Forecast step: Filtering dist. at time $t-1$ $\pi_{t-1\,|\,t-1} \rightarrow$ Forecast dist. $\pi_{t\,|\,t-1}$
   We obtain samples $\{x^{(1)}, \dots, x^{(M)}\} \sim \pi_{t\,|\,t-1}$

2. Analysis step: Forecast dist. $\pi_{t\,|\,t-1} \rightarrow$ Filtering dist. at time t $\pi_{t\,|\,t}$
   We obtain samples $\{x^{(1)}, \dots, x^{(M)}\} \sim \pi_{t\,|\,t}$

Ensemble filtering algorithms share the same forecast step but differ in the analysis step.

A “transformative” view of the analysis step



Analysis step: application of the **analysis map** T: Prior $\pi_{t|t-1} \to$ Posterior $\pi_{t|t}$

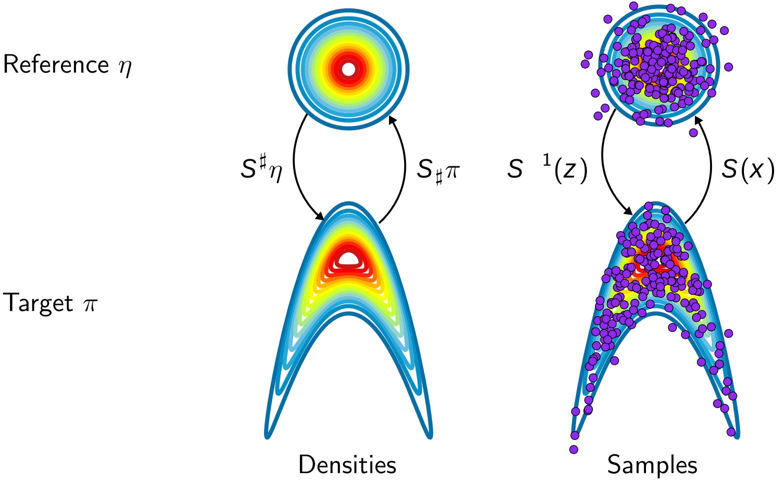
The analysis map of the Kalman filter T_{KF}:

$$T_{KF}(y, x) = x - \boldsymbol{\Sigma}_{\mathbf{X},\mathbf{Y}} \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}(y - y^\star) = x - K(y - y^\star)$$

The ensemble Kalman filter (EnKF) [Evensen, 1994] constructs an estimate $\widehat{K} \in \mathbb{R}^{n \times d}$ from limited samples $\{x^1, \dots, x^M\}$ of the forecast distribution.

Transport map between two probability measures

- Seek a transport map S that pushes forward π to η, i.e. $S_\sharp \pi = \eta$.
- Generate cheap and independent samples $x \quad \pi$) $S(x) \quad \eta$.



Looking for a good map [Marzouk et al., 2016]

Consider the **Knothe-Rosenblatt (KR) rearrangement** S s.t. $S_\sharp \pi = \eta$

$$S(z) = S(z_1, z_2, \ldots, z_m) = \begin{bmatrix} S^1(z_1) \\ S^2(z_1, z_2) \\ \vdots \\ S^m(z_1, z_2, \ldots, z_m) \end{bmatrix}.$$

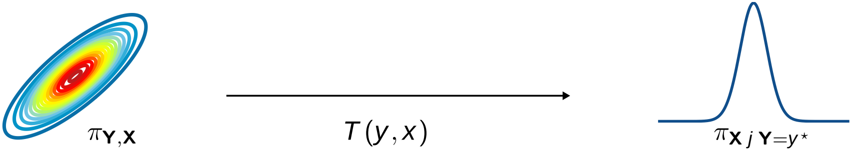
The KR has many nice features for Bayesian inference, e.g. easily invertible and $\det \nabla S(x)$ is simple to evaluate [Marzouk et al., 2016, Baptista et al., 2020].

The 1D map $\xi \mapsto S^k(x_1, x_2, \ldots, x_{k-1}, \xi)$ characterizes the marginal conditional $\pi_{\mathsf{X}_k | \mathbf{X}_{1:k-1} = x_{1:k-1}}(\xi)$.

Gaussian case

Consider $\mathbf{X} \sim \pi_{\mathbf{X}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $LL^\top = \boldsymbol{\Sigma}^{-1}$ be the Cholesky factorization of $\boldsymbol{\Sigma}^{-1}$.
Then $S(x) = L(x - \boldsymbol{\mu})$ is the KR that pushes forward $\pi_{\mathbf{X}}$ to $\eta = \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$.

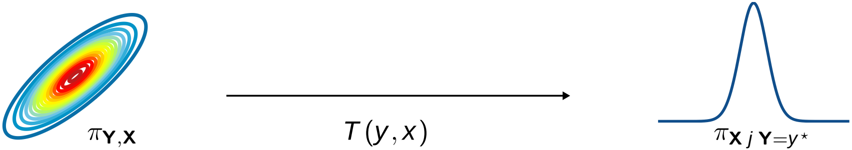
Construction of the analysis map [Spantini et al., 2022]



Construction of the analysis map [Spantini et al., 2022]

Consider the KR rearrangement S s.t. $S_\sharp \pi_{\mathbf{Y},\mathbf{X}} = \eta$

$$S(y, x) = \begin{bmatrix} S^{\mathcal{Y}}(y) \\ S^{\mathcal{X}}(y, x) \end{bmatrix},$$

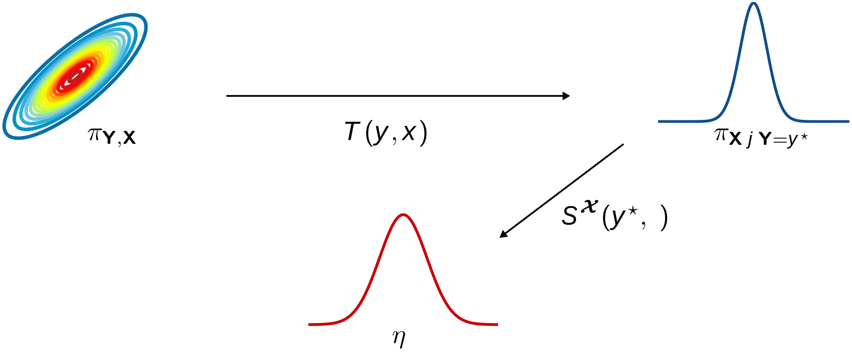


Construction of the analysis map [Spantini et al., 2022]

Consider the KR rearrangement S s.t. $S_\sharp \pi_{\mathbf{Y},\mathbf{X}} = \eta$

$$S(y, x) = \begin{bmatrix} S^{\mathcal{Y}}(y) \\ S^{\mathcal{X}}(y, x) \end{bmatrix},$$

The map $\xi \mapsto S^{\mathcal{X}}(y^\star, \xi)$ pushes forward $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|y^\star)$ to η



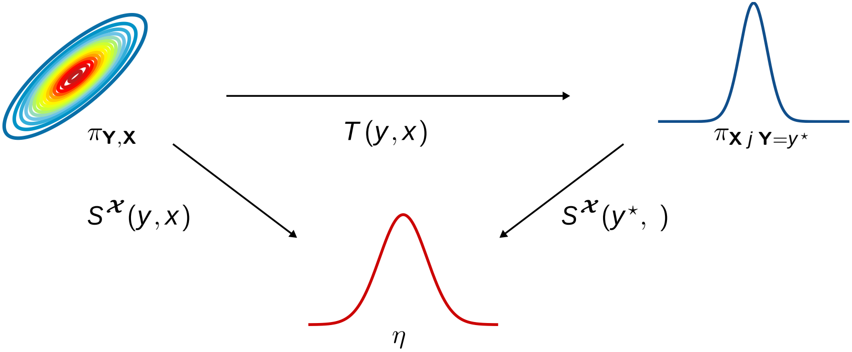
Construction of the analysis map [Spantini et al., 2022]

Consider the KR rearrangement S s.t. $S_\sharp \pi_{\mathbf{Y},\mathbf{X}} = \eta$

$$S(y, x) = \begin{bmatrix} S^{\mathcal{Y}}(y) \\ S^{\mathcal{X}}(y, x) \end{bmatrix},$$

The map $\xi \mapsto S^{\mathcal{X}}(y^\star, \xi)$ pushes forward $\pi_{\mathbf{X} | \mathbf{Y}}(\cdot | y^\star)$ to η

$S^{\mathcal{X}}(\mathbf{Y}, \mathbf{X}) \sim \eta$



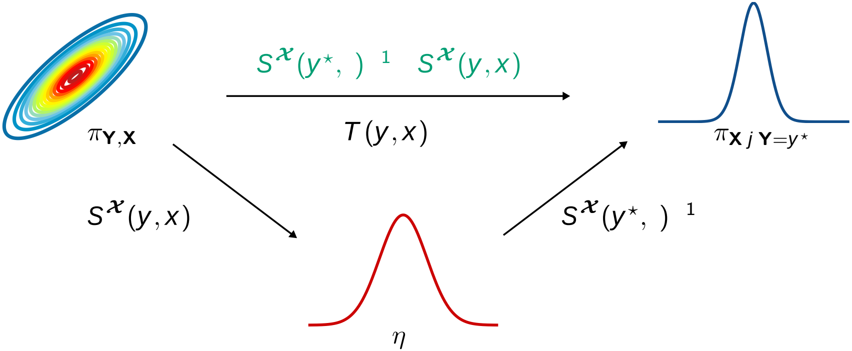
Construction of the analysis map [Spantini et al., 2022]

Consider the KR rearrangement S s.t. $S_\sharp \pi_{\mathbf{Y},\mathbf{X}} = \eta$

$$S(y,x) = \begin{bmatrix} S^{\mathcal{Y}}(y) \\ S^{\mathcal{X}}(y,x) \end{bmatrix},$$

The map $\xi \mapsto S^{\mathcal{X}}(y^\star, \xi)$ pushes forward $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|y^\star)$ to η

$S^{\mathcal{X}}(\mathbf{Y},\mathbf{X}) \sim \eta$



Derivation of the analysis map of the Kalman filter [Spantini et al., 2022]

$$T(y, x) = S^{\boldsymbol{\mathcal{X}}}(y^\star, \ )^{\ 1} \quad S^{\boldsymbol{\mathcal{X}}}(y, x)$$

We recover the analysis map of the Kalman filter when

1. $S^{\boldsymbol{\mathcal{X}}}$ is linear
2. The reference density is the standard normal distribution $\eta = N(\mathbf{0}_{d+n}, I_{d+n})$

Ensemble filters differ in the choice of

- the reference density
- the class of functions to represent $S^{\boldsymbol{\mathcal{X}}}$
- the estimation of $S^{\boldsymbol{\mathcal{X}}}$ from samples

Limitations of light-tailed filter for heavy-tailed distributions

Many ensemble filters like the EnKF assume **at least** that the **tails of the forecast distribution are Gaussian**.

These filters don't provide consistent inference for heavy-tailed filtering problems.

Our contribution: Introduce a new ensemble filter called **ensemble robust filter (EnRF)** based on the following assumptions:

- We restrict $S^{\boldsymbol{x}}$ to be linear
- We choose a reference distribution whose **tail-heaviness can be adapted to the data**.

t–distributions

t–distributions are a family of distributions parameterized by a mean $\boldsymbol{\mu}_{\mathbf{X}} \in \mathbb{R}^n$, a scale matrix $C_{\mathbf{X}} \in \mathbb{R}^{n \times n}$, and a degree of freedom $\nu_{\mathbf{X}} \in [1, \infty[$.

The degree of freedom $\nu_{\mathbf{X}}$ characterizes the tail-heaviness:

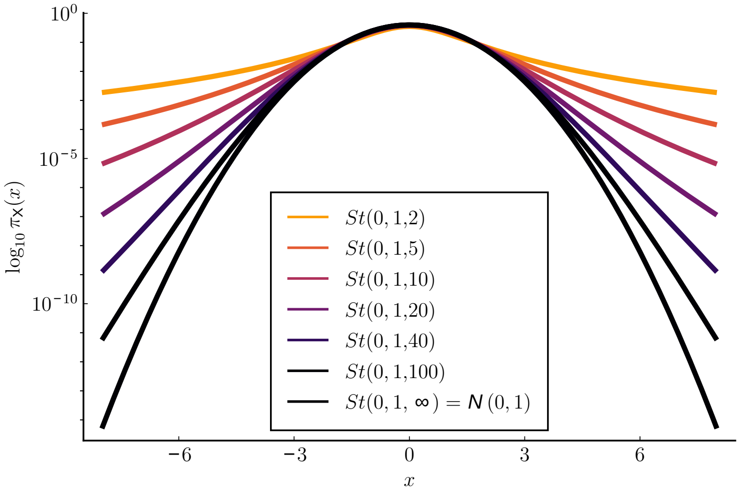
- For $\nu_{\mathbf{X}} = 1$, we recover the Cauchy distribution
- For $\nu_{\mathbf{X}} = \infty$, we recover the Gaussian distribution

Relation between the parameters and the moments of a t–distribution:

$$\mathrm{E}_{\pi_{\mathbf{X}}}[x] = \boldsymbol{\mu}_{\mathbf{X}}, \text{ for } \nu_{\mathbf{X}} > 1$$

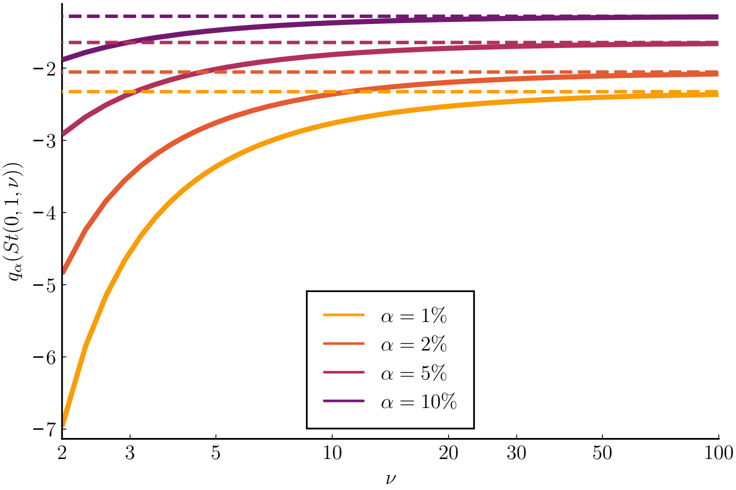
$$\mathrm{E}_{\pi_{\mathbf{X}}}\left[(x - \boldsymbol{\mu}_{\mathbf{X}})(x - \boldsymbol{\mu}_{\mathbf{X}})^\top\right] = \frac{\nu_{\mathbf{X}}}{\nu_{\mathbf{X}} - 2} C_{\mathbf{X}} \text{ for } \nu_{\mathbf{X}} > 2.$$

Probability density function of t–distributions



PDF of the univariate standard t–distribution $St\,(0,1,\nu)$ for $\nu = 2,5,10,20,100$ and the univariate standard Gaussian distribution $N\,(0,1) = St\,(0,1,\ 1\ )$ (black).

Quantiles of the t–distributions



Evolution of α-quantile q_α for $\alpha = 1\%, 2\%, 5\%, 10\%$ with the degree of freedom ν of the univariate standard t–distribution. Dashed lines corresponds to the α-quantiles for the standard Gaussian distribution $N(0,1) = St(0,1, \mathit{1})$.

Useful properties of t–distributions:

1. t–distributions are closed under affine transformations:

   If $\mathbf{X} \sim St\left(\boldsymbol{\mu}_{\mathbf{X}}, C_{\mathbf{X}}, \nu_{\mathbf{X}}\right)$, then $\mathbf{Z} = A\mathbf{X} + b \sim St\left(A\boldsymbol{\mu}_{\mathbf{X}} + b, AC_{\mathbf{X}}A^\top, \nu_{\mathbf{X}}\right)$

2. Conditional and marginal distributions are known in closed form.

   Consider $\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix} \sim St\left(\begin{bmatrix}\boldsymbol{\mu}_{\mathbf{X}}\\ \boldsymbol{\mu}_{\mathbf{Y}}\end{bmatrix}, \begin{bmatrix}C_{\mathbf{Y}} & C_{\mathbf{X},\mathbf{Y}}^\top\\ C_{\mathbf{X},\mathbf{Y}} & C_{\mathbf{X}}\end{bmatrix}, \nu\right)$

   Marginal: $\mathbf{X} \sim St\left(\boldsymbol{\mu}_{\mathbf{X}}, C_{\mathbf{X}}, \nu_{\mathbf{X}}\right)$

   Conditional: $\pi_{\mathbf{X}\mid\mathbf{Y}=y} \sim St\left(\boldsymbol{\mu}_{\mathbf{X}\mid\mathbf{Y}=y}, C_{\mathbf{X}\mid\mathbf{Y}=y}, \nu_{\mathbf{X}\mid\mathbf{Y}=y}\right)$ with

$$\boldsymbol{\mu}_{\mathbf{X}\mid y} = \boldsymbol{\mu}_{\mathbf{X}} + C_{\mathbf{X},\mathbf{Y}}C_{\mathbf{Y}}^{-1}(y - \boldsymbol{\mu}_{\mathbf{Y}})$$

$$C_{\mathbf{X}\mid\mathbf{Y}=y} = \underbrace{\frac{\nu + (y - \boldsymbol{\mu}_{\mathbf{Y}})^\top C_{\mathbf{Y}}^{-1}(y - \boldsymbol{\mu}_{\mathbf{Y}})}{\nu + d}}_{\alpha_{\mathbf{Y}}(y) > 0}\underbrace{\left(C_{\mathbf{X}} - C_{\mathbf{X},\mathbf{Y}}C_{\mathbf{Y}}^{-1}C_{\mathbf{X},\mathbf{Y}}^\top\right)}_{\text{Schur complement } C_{\mathbf{X}\backslash\mathbf{Y}}}$$

$$\nu_{\mathbf{X}\mid\mathbf{Y}=y} = \nu + d$$

Analysis map T_ν for t–distributions

Let S_ν the KR that pushes forward the joint t–distribution $\pi_{\mathbf{Y},\mathbf{X}}$ with dof ν to a "judicious" t–distribution η_ν with same dof, i.e. $S_{\nu\sharp}\pi_{\mathbf{Y},\mathbf{X}} = \eta_\nu$.

Key: S_ν can be computed in closed form (new result).

We obtain the analysis map $T_\nu : \mathrm{R}^d \quad \mathrm{R}^n \; ! \; \mathrm{R}^n$ by partial inversion of $S_\nu^{\boldsymbol{\mathcal{X}}}$:

$$
\begin{aligned}
T_\nu(y, x) &= S_\nu^{\boldsymbol{\mathcal{X}}}(y^\star, \;)^{\;\;1} \quad S_\nu^{\boldsymbol{\mathcal{X}}}(y, x) \\
&= \boldsymbol{\mu}_{\mathbf{X}} + C_{\mathbf{X},\mathbf{Y}} C_{\mathbf{Y}}^{\;\;1}(y^\star \quad \boldsymbol{\mu}_{\mathbf{Y}}) + \sqrt{\frac{\alpha_{\mathbf{Y}}(y^\star)}{\alpha_{\mathbf{Y}}(y)}} \left[(x \quad \boldsymbol{\mu}_{\mathbf{X}}) \quad C_{\mathbf{X},\mathbf{Y}} C_{\mathbf{Y}}^{\;\;1}(y \quad \boldsymbol{\mu}_{\mathbf{Y}})\right]
\end{aligned}
$$

Note that $C_{\mathbf{X},\mathbf{Y}} C_{\mathbf{Y}}^{\;\;1} = \boldsymbol{\Sigma}_{\mathbf{X},\mathbf{Y}} \boldsymbol{\Sigma}_{\mathbf{Y}}^{\;\;1}$.

Connection with the Kalman filter

Let's perform an asymptotic expansion of T_ν for large ν:

$$T_\nu = \underbrace{T_1}_{T_{KF}} + \text{higher order terms in } O\left(\frac{1}{\nu}\right)$$

The zeroth order term T_1 is exactly the analysis map of the Kalman filter T_{KF}.

The higher order terms correct the analysis map T_{KF} to account for the finite degree of freedom.

Takeaway

The analysis map T_ν generalizes the analysis map of the Kalman filter for t–distributions with finite degree of freedom!

Sensitivity to outlying synthetic observations

Context: Observation models often rely on simplified physics and suffer from mis-specifications of the observation operator and observation noise.

Let $(y^{(j)}, x^{(j)})$ be a joint forecast sample with an outlying synthetic observation $y^{(j)}$ generated by the likelihood model $\pi_{\mathbf{Y} \mid \mathbf{X}=x^{(j)}}$, such that $\delta_{\mathbf{Y}}\big(y^{(j)}\big) = \big(y^{(j)} - \boldsymbol{\mu}_{\mathbf{Y}}\big)^{\top} C_{\mathbf{Y}}^{-1} \big(y^{(j)} - \boldsymbol{\mu}_{\mathbf{Y}}\big) \to \infty$. Then, the analysis map T_{ν} reduces to

$$T_{\nu}(y, x) = \boldsymbol{\mu}_{\mathbf{X}} + C_{\mathbf{X},\mathbf{Y}} C_{\mathbf{Y}}^{-1}(y^{\star} - \boldsymbol{\mu}_{\mathbf{Y}}).$$

Takeaways:

- The outlying observation $y^{(j)}$ vanishes from the analysis map.
- The prior sample $x^{(j)}$ is mapped to the posterior mean/median/mode.

What is the pushforward of $\pi_{\mathbf{X}}$ by the maps T_{ν} and T_{KF}?

T_{ν} leads to exact inference: $T_{\nu\sharp}\pi_{\mathbf{X}} = \pi_{\mathbf{X}|y^\star} = St\left(\boldsymbol{\mu}_{\mathbf{X}|y^\star}, C_{\mathbf{X}|y^\star}, \nu_{\mathbf{X}|y^\star}\right)$:

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{X}|y^\star} &= \boldsymbol{\mu}_{\mathbf{X}} + C_{\mathbf{X},\mathbf{Y}} C_{\mathbf{Y}}^{-1}(y^\star - \boldsymbol{\mu}_{\mathbf{Y}}), \\ C_{\mathbf{X}|y^\star} &= \alpha_{\mathbf{Y}}(y^\star) C_{\mathbf{X}\backslash\mathbf{Y}}, \\ \nu_{\mathbf{X}|y^\star} &= \nu + d, \end{aligned} \tag{1}$$

We interpret $\alpha_{\mathbf{Y}}(y^\star)$ as an **adaptive** and **data-dependent** multiplicative inflation.

$T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}} = St\left(\boldsymbol{\mu}_{T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}}}, C_{T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}}}, \nu_{T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}}}\right)$ with

$$\begin{aligned} \boldsymbol{\mu}_{T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}}} &= \boldsymbol{\mu}_{\mathbf{X}} + C_{\mathbf{X},\mathbf{Y}} C_{\mathbf{Y}}^{-1}(y^\star - \boldsymbol{\mu}_{\mathbf{Y}}), \\ C_{T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}}} &= 1 \times C_{\mathbf{X}\backslash\mathbf{Y}}, \\ \nu_{T_{\mathrm{KF}\sharp}\pi_{\mathbf{X}}} &= \nu + 0. \end{aligned} \tag{2}$$

The Kalman filter is only consistent to estimate the mean value of t–distributions.

Challenges of estimating heavy-tailed distributions from samples

The classical sample mean and covariance estimators are derived from a maximum likelihood approach for Gaussian distributions.

These light-tailed estimators are very sensitive to outliers and introduce additional variance.

We use a "regularized" expectation-maximization algorithm (`EMq`) to estimate the heavy-tailed joint forecast distribution $\pi_{\mathbf{Y},\mathbf{X}}$ from $f(y^{(i)}, x^{(i)})g$ [Doğru et al., 2018].

We call **ensemble robust filter (EnRF)** the ensemble filter that estimates T_{ν} with the `EMq` from the joint forecast samples $f(y^{(i)}, x^{(i)})g$.

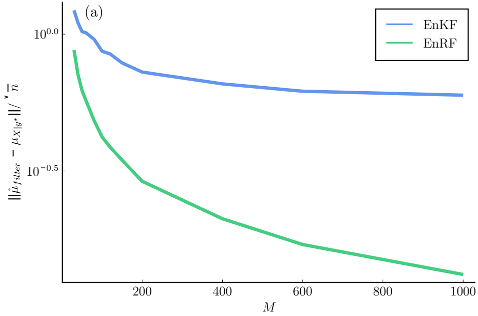
Empirical performance of light-tailed and heavy-tailed estimators

Consider $\begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \quad St\left(\begin{bmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{bmatrix}, \begin{bmatrix} C_{\mathbf{Y}} & C_{\mathbf{X},\mathbf{Y}}^{>} \\ C_{\mathbf{X},\mathbf{Y}} & C_{\mathbf{X}} \end{bmatrix}, \nu\right)$ with $\nu = 2.1$, $\mathbf{X}$ *2* R^{10} and $\mathbf{Y}$ *2* R^{5}.

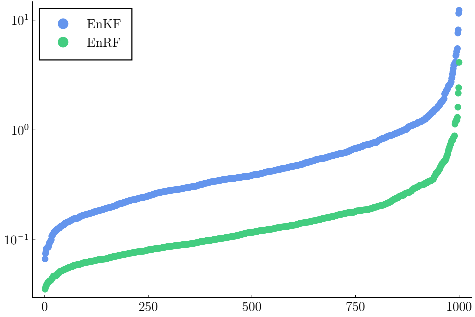
Experiment:

1. We generate M samples $f(y^{(i)}, x^{(i)})g \quad \pi_{\mathbf{Y},\mathbf{X}}$.
2. We apply the analysis map of the EnKF and the EnRF to assimilate a realization $y^{\star} \quad \pi_{\mathbf{Y}}$
3. We compute the sample mean and covariance for the two posterior ensembles.

Posterior mean estimates



Left panel: Evolution of the error RMSE $= ||\widehat{\boldsymbol{\mu}}_{filter} - \boldsymbol{\mu}_{\mathbf{X}|y^\star}||_2/\sqrt{n}$ with the ensemble size M.

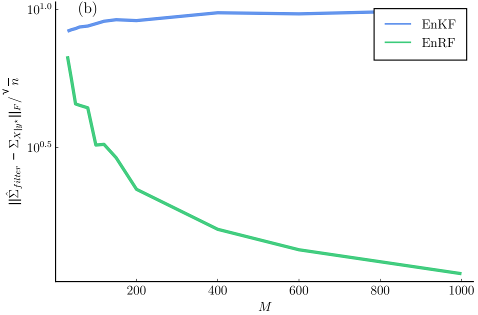


Right panel: Empirical distribution of the RMSE over 1000 realizations of $y^\star$ using $M = 600$ samples.

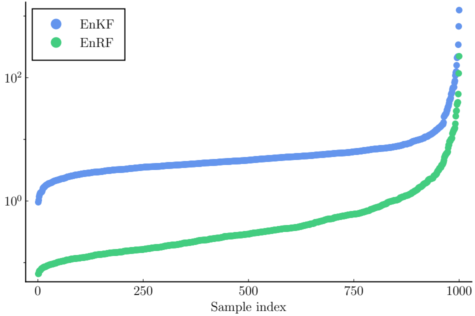
Both filters are consistent,

but the sEnKF has a much slower convergence rate than the EnRF.

Posterior covariance estimates



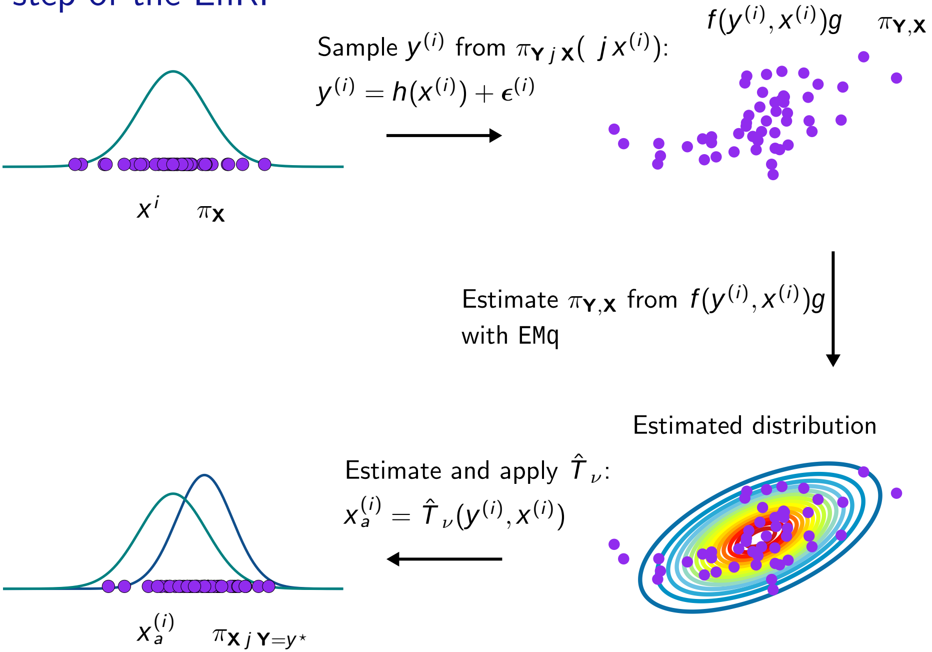
Left panel: Evolution of the error $||\widehat{\boldsymbol{\Sigma}}_{filter} - \boldsymbol{\Sigma}_{\mathbf{X}|y^\star}||_F/\sqrt{n}$ with the ensemble size M.



Right panel: Empirical distribution of the error $||\widehat{\boldsymbol{\Sigma}}_{filter} - \boldsymbol{\Sigma}_{\mathbf{X}|y^\star}||_F/\sqrt{n}$ over 1000 realizations of $y^\star$ using $M = 600$.

The sEnKF is not consistent to estimate covariances of t–distributions.

Analysis step of the EnRF



Estimate the tail-heaviness on the fly

We can estimate the dof at each assimilation cycle with the `EMq`

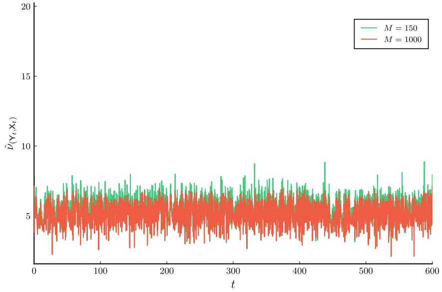
The EnRF can adapt its analysis map to the tail heaviness of the data!

Limitation: Computational cost of the `EMq` when the dof is unknown.

Idea: Present 3 variants of the EnRF that differ in the frequency and the samples used to estimate the dof

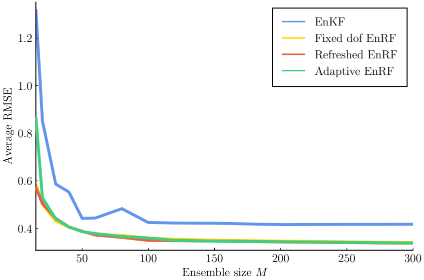
1. The `RefreshEnRF`: We maintain a buffer of past filtering samples, and estimate the dof with this larger ensemble every $\Delta t_{refresh}$.
2. The `FixedEnRF`: The dof is estimated once from a free-run of the state-space model and fixed for the entire assimilation ($\Delta t_{refresh} = \mathit{1}$)
3. The `AdaptEnRF`: The dof is estimated at each assimilation cycle ($\Delta t_{refresh} = \Delta t_{obs}$)

Lorenz-63



Empirical degree of freedom $\hat{\nu}_{(\mathbf{Y}_t,\mathbf{X}_t)}$ with $M = 150, 1000$ samples of the joint forecast density $\{(\boldsymbol{y}_t^{(i)}, \boldsymbol{x}_t^{(i)})\} \sim \pi_{(\mathbf{Y}_t,\mathbf{X}_t) | \mathbf{Y}_{1:t-1}}$ for the Lorenz-63 problem with t–distributed observation noise.

RMSE results for Lorenz-63



sEnKF FixedEnRF RefreshEnRF AdaptEnRF

Evolution of the RMSE with the ensemble size M for the Lorenz-63 model with t–distributed observation noise with $\nu = 3.0$.

We optimally tune the multiplicative inflation of the sEnKF.

25% reduction of the RMSE with the EnRF!

The different EnRFs don't require tuning: Plug and Play!

Conclusion and Outlook

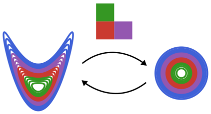
Summary:

We introduce the EnRF that generalizes the EnKF to heavy-tailed t–distributions

The EnRF adapts its prior-to-posterior to the tail-heaviness of the data
$\Longrightarrow$ Adaptive and data-dependent multiplicative inflation

The EnRF requires no tuning: **Plug and Play!**

Software: The EnRF will be soon available in `TransportBasedInference.jl`



Acknowledgements: Support from National Science Foundation is gratefully acknowledged.

Contact: mathieuleprovost1@gmail.com

References I

Baptista, R., Zahm, O., and Marzouk, Y. (2020).
An adaptive transport framework for joint and conditional density estimation.
arXiv preprint arXiv:2009.10303.

Doğru, F. Z., Bulut, Y. M., and Arslan, O. (2018).
Doubly reweighted estimators for the parameters of the multivariate t-distribution.
Communications in Statistics-Theory and Methods, 47(19):4751–4771.

Evensen, G. (1994).
Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics.
Journal of Geophysical Research: Oceans, 99(C5):10143–10162.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016).
Sampling via measure transport: An introduction.
Handbook of Uncertainty Quantification, pages 1–41.

Spantini, A., Baptista, R., and Marzouk, Y. (2022).
Coupling Techniques for Nonlinear Ensemble Filtering.
SIAM Review, 64(4):921–953.
Publisher: Society for Industrial and Applied Mathematics.