Nonlinear and/or Non-normal Filtering

Jesús Fernández-Villaverde
University of Pennsylvania

Motivation

- Nonlinear and/or non-gaussian filtering, smoothing, and forecasting (NLGF) problems are pervasive in economics.
- Macroeconomics: evaluating likelihood of DSGE models
- Finance: time-varying variance of time series.
- However, NLGF is a complicated endeavor with no simple and exact algorithm.

Environment

- Discrete time $t \in \{1, 2, ...\}$.
- States S_t.
- Initial state S_0 is either known or it comes from $p(S_0; \gamma)$.
- Properties of $p(S_0; \gamma)$? Stationarity?

Nonlinear and/or Non-gaussian State Space Representations

- Transition equation:

$$S_t = f\left(S_{t-1}, W_t; \gamma\right)$$

- Measurement equation:

$$Y_t = g\left(S_t, V_t; \gamma\right)$$

- Interpretation.

Shocks

- $\{W_t\}$ and $\{V_t\}$ are independent of each other.
- W_t and V_t have zero mean.
- The variance of W_t is given by R_t and the variance of V_t by Q_t.
- Generalizations?

Conditional Densities

- From $S_t = f\left(S_{t-1}, W_t; \gamma\right)$, we can compute $p\left(S_t | S_{t-1}; \gamma\right)$.

- From $Y_t = g\left(S_t, V_t; \gamma\right)$, we can compute $p\left(Y_t | S_t; \gamma\right)$.

- From $S_t = f\left(S_{t-1}, W_t; \gamma\right)$ and $Y_t = g\left(S_t, V_t; \gamma\right)$, we have:

$$Y_t = g\left(f\left(S_{t-1}, W_t; \gamma\right), V_t; \gamma\right)$$

  and hence we can compute $p\left(Y_t | S_{t-1}; \gamma\right)$.

Filtering, Smoothing, and Filtering

- Filtering: we are concerned with what we have learned up to current observation.
- Smoothing: we are concerned with what we learn with the full sample.
- Forecasting: we are concerned with future realizations.

Goal of Filtering I

- Compute conditional densities: $p\left(s_t|y^{t-1};\gamma\right)$ and $p\left(s_t|y^{t};\gamma\right)$.

- Why?

  1. It allows probability statements regarding the situation of the system.

  2. Compute conditional moments: mean, $s_{t|t}$ and $s_{t|t-1}$ and variances $P_{t|t}$ and $P_{t|t-1}$.

  3. Other functions of the states.

Goals of Filtering II

- Compute condition the likelihood function of a sequence of realizations of the observable y^T at a particular parameter value γ:

$$p\left(y^T;\gamma\right)$$

- Given the markov structure of our state space representation, we can factorize the likelihood function as"

$$p\left(y^T;\gamma\right) = \prod_{t=1}^{T} p\left(y_t|y^{t-1};\gamma\right)$$

Goals of Filtering III

- Then,

$$\begin{aligned} p\left(y^T;\gamma\right) &= p\left(y_1|\gamma\right)\prod_{t=1}^{T} p\left(y_t|y^{t-1};\gamma\right) \\ &= \int p\left(y_1|s_0;\gamma\right)ds_0 \prod_{t=1}^{T}\int p\left(y_t|s_t;\gamma\right)p\left(s_t|y^{t-1};\gamma\right)ds_t \end{aligned}$$

- Hence, knowledge of $\left\{p\left(s_t|y^{t-1};\gamma\right)\right\}_{t=1}^{T}$ and $p\left(s_0;\gamma\right)$ allow the evaluation of the likelihood of the model.

Two Fundamental Tools

1. Chapman-Kolmogorov equation:

$$p\left(s_t|y^{t-1};\gamma\right) = \int p\left(s_t|s_{t-1};\gamma\right) p\left(s_{t-1}|y^{t-1};\gamma\right) ds_{t-1}$$

2. Bayes' theorem:

$$p\left(s_t|y^{t};\gamma\right) = \frac{p\left(y_t|s_t;\gamma\right) p\left(s_t|y^{t-1};\gamma\right)}{p\left(y_t|y^{t-1};\gamma\right)}$$

where:

$$p\left(y_t|y^{t-1};\gamma\right) = \int p\left(y_t|s_t;\gamma\right) p\left(s_t|y^{t-1};\gamma\right) ds_t$$

Interpretation

1. Chapman-Kolmogorov equation is one-step ahead predictor.

2. Bayes' theorem updates the conditional density of states given the new observation.

Recursion for $p\left(s_t|y^t;\gamma\right)$

- Combining the Chapman-Kolmogorov and the Bayes' theorem:

$$p\left(s_t|y^t;\gamma\right) = \frac{\int p\left(s_t|s_{t-1};\gamma\right) p\left(s_{t-1}|y^{t-1};\gamma\right) ds_{t-1}}{\int \left\{\int p\left(s_t|s_{t-1};\gamma\right) p\left(s_{t-1}|y^{t-1};\gamma\right) ds_{t-1}\right\} p\left(y_t|s_t;\gamma\right) ds_t} p\left(y_t|s_t;\gamma\right)$$

- To initiate that recursion, we only need a value for s_0 or $p\left(s_0;\gamma\right)$.

- Applying the Chapman-Kolmogorov equation once more to the outcome of the recursion, the get the sequence of $\left\{p\left(s_t|y^{t-1};\gamma\right)\right\}_{t=1}^{T}$ to evaluate the likelihood function.

Smoothing

- We are interested on the distribution of the state conditional on all the observations, on $p\left(s_t|y^T;\gamma\right)$ and $p\left(y_t|y^T;\gamma\right)$.

- We compute:

$$p\left(s_t|y^T;\gamma\right) = p\left(s_t|y^t;\gamma\right)\int \frac{p\left(s_{t+1}|y^T;\gamma\right)p\left(s_{t+1}|s_t;\gamma\right)}{p\left(s_{t+1}|y^t;\gamma\right)}ds_{t+1}$$

  a backward recursion that we initialize with $p\left(s_T|y^T;\gamma\right)$ and the sequences $\left\{p\left(s_t|y^t;\gamma\right)\right\}_{t=1}^T$ and $\left\{p\left(s_t|y^{t-1};\gamma\right)\right\}_{t=1}^T$ we obtained from filtering.

Forecasting

- We apply the Chapman-Kolmogorov equation recursively, we can get $p\left(s_{t+j}|y^t;\gamma\right),\ j \geq 1$.

- Integrating recursively:

$$p\left(y_{l+1}|y^l;\gamma\right) = \int p\left(y_{l+1}|s_{l+1};\gamma\right) p\left(s_{l+1}|y^l;\gamma\right) ds_{l+1}$$

  from $t+1$ to $t+j$, we get $p\left(y_{t+j}|y^T;\gamma\right)$.

- Clearly smoothing and forecasting require to solve the filtering problem first!

Problem of Filtering

- We have the recursion

$$p\left(s_t|y^t;\gamma\right)=\frac{\int p\left(s_t|s_{t-1};\gamma\right)p\left(s_{t-1}|y^{t-1};\gamma\right)ds_{t-1}}{\int\left\{\int p\left(s_t|s_{t-1};\gamma\right)p\left(s_{t-1}|y^{t-1};\gamma\right)ds_{t-1}\right\}p\left(y_t|s_t;\gamma\right)ds_t}p\left(y_t|s_t;\gamma\right)$$

- A lot of complicated and high dimensional integrals.

- In general, we do not have close for solution for them.

- Nonlinear and non-gaussianity translate, spread, and deform (TSD) the conditional densities in ways that impossibilities to accommodate them with any known parametric family.

Exception

- There is one exception: linear and gaussian case.
- Why? Because if the system is linear and gaussian, all the conditional probabilities are also gaussian.
- Linear and gaussian state spaces models translate and spread the conditional distributions, but they do not deform them.
- For gaussian distributions, we only need to track mean and variance (sufficient statistics).
- Kalman filter accomplishes this goal efficiently.

Different Approaches

- Deterministic filtering:
  1. Kalman family.
  2. Grid-based filtering.

- Simulation filtering:
  1. McMc.
  2. Sequential Monte Carlo.

Kalman Family of Filters

- Use ideas of Kalman filtering to NLGF problems.
- Non-optimal filters.
- Different implementations:
  1. Extended Kalman filter.
  2. Iterated Extended Kalman filter.
  3. Second-order Extended Kalman filter.
  4. Unscented Kalman filter.

The Extended Kalman Filter

- EKF is historically the first descendant of the Kalman filter.
- EKF deals with nonlinearities with a first order approximation to the system and applying the Kalman filter to this approximation.
- Non-gaussianities are ignored.

Algorithm I

- Given $s_{t-1|t-1}$:

$$s_{t|t-1} = f\left(s_{t-1|t-1}, 0; \gamma\right)$$

- Then:

$$P_{t|t-1} = Q_{t-1} + F_t P_{t-1|t-1} F_t'$$

where

$$F_t = \left.\frac{df\left(S_{t-1}, W_t; \gamma\right)}{dS_{t-1}}\right|_{S_{t-1}=s_{t-1|t-1}, W_t=0}$$

Algorithm II

- Kalman gain, K_t, is:

$$K_t = P_{t|t-1} G_t' \left(G_t P_{t|t-1} G_t' + R_t \right)^{-1}$$

where

$$G_t = \left. \frac{dg\left(S_{t-1}, v_t; \gamma\right)}{dS_{t-1}} \right|_{S_{t-1}=s_{t|t-1}, v_t=0}$$

- Then

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + K_t \left(y_t - g\left(s_{t|t-1}, 0; \gamma \right) \right) \\ P_{t|t} &= P_{t|t-1} - K_t G_t P_{t|t-1} \end{aligned}$$

Problems of EKF

1. It ignores the non-gaussianities of W_t and V_t.

2. It ignores the non-gaussianities of states distribution.

3. Approximation error incurred by the linearization.

4. Biased estimate of the mean and variance.

5. We need to compute Jacobian and Hessians.

As the sample size grows, those errors accumulate and the filter diverges.

Iterated Extended Kalman Filter I

- Compute $s_{t|t-1}$ and $P_{t|t-1}$ as in EKF.

- Iterate N times on:

$$K_t^i = P_{t|t-1} G_t^{i\prime} \left(G_t^i P_{t|t-1} G_t^{i\prime} + R_t \right)^{-1}$$

where

$$G_t^i = \left. \frac{dg\left(S_{t-1}, v_t; \gamma\right)}{dS_{t-1}} \right|_{S_{t-1}=s_{t|t-1}^i, v_t=0}$$

and

$$s_{t|t}^i = s_{t|t-1} + K_t^i \left(y_t - g\left(s_{t|t-1}, 0; \gamma \right) \right)$$

Iterated Extended Kalman Filter II

- Why are we iterating?
- How many times?
- Then:

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + K_t \left(y_t - g\left(s^N_{t|t-1}, 0; \gamma \right) \right) \\ P_{t|t} &= P_{t|t-1} - K^N_t G^N_t P_{t|t-1} \end{aligned}$$

Second-order Extended Kalman Filter

- We keep second-order terms of the Taylor expansion of transition and measurement.
- Theoretically, less biased than EKF.
- Messy algebra.
- In practice, not much improvement.

Unscented Kalman Filter I

- Recent proposal by Julier and Uhlmann (1996).
- Based around the unscented transform.
- A set of sigma points is selected to preserve some properties of the conditional distribution (for example, the first two moments).
- Then, those points are transformed and the properties of the new conditional distribution are computed.

Unscented Kalman Filter II

- The UKF computes the conditional mean and variance accurately up to a third order approximation if the shocks W_t and V_t are gaussian and up to a second order if they are not.

- The sigma points are chosen deterministically and not by simulation as in a Monte Carlo method.

- The UKF has the advantage with respect to the EKF that no jacobian or hessians is required, objects that may be difficult to compute.

New State Variable

- We modify the state space by creating a new augmented state variable:

$$\mathbb{S}_t = [S_t, W_t, V_t]$$

that includes the pure state space and the two random variables W_t and V_t.

- We initialize the filter with

$$\mathbf{s}_{0|0} = E\left(\mathbb{S}_t\right) = E\left(S_0, 0, 0\right)$$

$$\mathbb{P}_{0|0} = \begin{bmatrix} P_{0|0} & 0 & 0 \\ 0 & R_0 & 0 \\ 0 & 0 & Q_0 \end{bmatrix}$$

Sigma Points

- Let L be the dimension of the state variable $\mathbb{S}_t$.

- For $t = 1$, we calculate the $2L + 1$ sigma points:

$$
\begin{aligned}
\mathcal{S}_{0,t-1|t-1} &= \mathbb{s}_{t-1|t-1} \\
\mathcal{S}_{i,t-1|t-1} &= \mathbb{s}_{t-1|t-1} - \left((L + \lambda) \, \mathbb{P}_{t-1|t-1} \right)^{0.5} \quad \text{for } i = 1, ..., L \\
\mathcal{S}_{i,t-1|t-1} &= \mathbb{s}_{t-1|t-1} + \left((L + \lambda) \, \mathbb{P}_{t-1|t-1} \right)^{0.5} \quad \text{for } i = L + 1, ..., 2L
\end{aligned}
$$

Parameters

- $\lambda = \alpha^2 (L + \kappa) - L$ is a scaling parameter.

- α determines the spread of the sigma point and it must belong to the unit interval.

- κ is a secondary parameter usually set equal to zero.

- Notation for each of the elements of $\mathcal{S}$:

$$\mathcal{S}_i = [\mathcal{S}_i^s, \mathcal{S}_i^w, \mathcal{S}_i^v] \text{ for } i = 0, ..., 2L$$

Weights

- Weights for each point:

$$\begin{aligned} \mathcal{W}_0^m &= \frac{\lambda}{L+\lambda} \\ \mathcal{W}_0^c &= \frac{\lambda}{L+\lambda} + \left(1 - \alpha^2 + \beta\right) \\ \mathcal{W}_0^m &= X_0^c = \frac{1}{2\left(L+\lambda\right)} \text{ for } i = 1, ..., 2L \end{aligned}$$

- β incorporates knowledge regarding the conditional distributions.

- For gaussian distributions, $\beta = 2$ is optimal.

Algorithm I: Prediction of States

- We compute the transition of the pure states:

$$\mathcal{S}_{i,t|t-1}^{s} = f\left(\mathcal{S}_{i,t|t-1}^{s}, \mathcal{S}_{i,t-1|t-1}^{w}; \gamma\right)$$

- Weighted state

$$s_{t|t-1} = \sum_{i=0}^{2L} \mathcal{W}_{i}^{m} \mathcal{S}_{i,t|t-1}^{s}$$

- Weighted variance:

$$P_{t|t-1} = \sum_{i=0}^{2L} \mathcal{W}_{i}^{c} \left(\mathcal{S}_{i,t|t-1}^{s} - s_{t|t-1}\right)\left(\mathcal{S}_{i,t|t-1}^{s} - s_{t|t-1}\right)'$$

Algorithm II: Prediction of Observables

- Predicted sigma observables:

$$\mathcal{Y}_{i,t|t-1} = g\left(\mathcal{S}^{s}_{i,t|t-1}, \mathcal{S}^{v}_{i,t|t-1}; \gamma\right)$$

- Predicted observable:

$$y_{t|t-1} = \sum_{i=0}^{2L} \mathcal{W}_i^m \mathcal{Y}_{i,t|t-1}$$

Algorithm III: Update

- Variance-covariance matrices:

$$
\begin{aligned}
P_{yy,t} &= \sum_{i=0}^{2L} \mathcal{W}_i^c \left(\mathcal{Y}_{i,t|t-1} - y_{t|t-1}\right)\left(\mathcal{Y}_{i,t|t-1} - y_{t|t-1}\right)' \\
P_{xy,t} &= \sum_{i=0}^{2L} \mathcal{W}_i^c \left(\mathcal{S}_{i,t|t-1}^s - s_{t|t-1}\right)\left(\mathcal{Y}_{i,t|t-1} - y_{t|t-1}\right)'
\end{aligned}
$$

- Kalman gain:

$$K_t = P_{xy,t} P_{yy,t}^{-1}$$

Algorithm IV: Update

- Update of the state:

$$s_{t|t} = s_{t|t} + K_t \left(y_t - y_{t|t-1} \right)$$

- the update of the variance:

$$P_{t|t} = P_{t|t-1} + K_t P_{yy,t} K_t'$$

Finally:

$$\mathbb{P}_{t|t} = \begin{bmatrix} P_{t|t} & 0 & 0 \\ 0 & R_t & 0 \\ 0 & 0 & Q_t \end{bmatrix}$$

Grid-Based Filtering

- Remember that we have the recursion

$$p\left(s_t|y^t;\gamma\right) = \frac{\int p\left(s_t|s_{t-1};\gamma\right) p\left(s_{t-1}|y^{t-1};\gamma\right) ds_{t-1}}{\int \left\{\int p\left(s_t|s_{t-1};\gamma\right) p\left(s_{t-1}|y^{t-1};\gamma\right) ds_{t-1}\right\} p\left(y_t|s_t;\gamma\right) ds_t} p\left(y_t|s_t;\gamma\right)$$

- This recursion requires the evaluation of three integrals.

- This suggests the possibility of addressing the problem by computing those integrals by a deterministic procedure as a grid method.

- Kitagawa (1987) and Kramer and Sorenson (1988).

Grid-Based Filtering

- We divide the state space into N cells, with center point s_t^i, $\left\{s_t^i : i = 1, ..., N\right\}$.
- We substitute the exact conditional densities by discrete densities that put all the mass at the points $\left\{s_t^i\right\}_{i=1}^N$.
- We denote $\delta(x)$ is a dirac function with mass at 0.

Approximated Densities and Weights

$$
\begin{aligned}
p\left(s_t|y^{t-1};\gamma\right) &\simeq \sum_{i=1}^{N} \omega_{t|t-1}^{i} \delta\left(s_t - s_t^i\right) \\
p\left(s_t|y^{t};\gamma\right) &\simeq \sum_{i=1}^{N} \omega_{t|t-1}^{i} \delta\left(s_t - s_t^i\right) \\
\omega_{t|t-1}^{i} &= \sum_{j=1}^{N} \omega_{t-1|t-1}^{j} p\left(s_t^i|s_{t-1}^j;\gamma\right) \\
\omega_{t|t}^{i} &= \frac{\omega_{t|t-1}^{i} p\left(y_t|s_t^i;\gamma\right)}{\sum_{j=1}^{N} \omega_{t|t-1}^{j} p\left(y_t|s_t^j;\gamma\right)}
\end{aligned}
$$

Approximated Recursion

$$p\left(s_t|y^t;\gamma\right)=\sum_{i=1}^{N}\frac{\left[\sum_{j=1}^{N}\omega_{t-1|t-1}^{j}p\left(s_t^i|s_{t-1}^j;\gamma\right)\right]p\left(y_t|s_t^i;\gamma\right)}{\sum_{j=1}^{N}\left[\sum_{j=1}^{N}\omega_{t-1|t-1}^{j}p\left(s_t^i|s_{t-1}^j;\gamma\right)\right]p\left(y_t|s_t^j;\gamma\right)}\delta\left(s_t-s_t^i\right)$$

Compare with

$$p\left(s_t|y^t;\gamma\right)=\frac{\int p\left(s_t|s_{t-1};\gamma\right)p\left(s_{t-1}|y^{t-1};\gamma\right)ds_{t-1}}{\int\left\{\int p\left(s_t|s_{t-1};\gamma\right)p\left(s_{t-1}|y^{t-1};\gamma\right)ds_{t-1}\right\}p\left(y_t|s_t;\gamma\right)ds_t}p\left(y_t|s_t;\gamma\right)$$

given that

$$p\left(s_{t-1}|y^{t-1};\gamma\right)\simeq\sum_{i=1}^{N}\omega_{t-1|t-1}^{i}\delta\left(s_t^i\right)$$

Problems

- Grid filters require a constant readjustment to small changes in the model or its parameter values.

- Too computationally expensive to be of any practical benefit beyond very low dimensions.

- Grid points are fixed ex ante and the results are very dependent on that choice.

Can we overcome those difficulties and preserve the idea of integration? Yes, through Monte Carlo Integration.