Pruned skewed Kalman filter and smoother with application to DSGE models

– Online appendix –

November 2025

Gaygysyz Guljanov[a], Willi Mutschler[b,], Mark Trede[a]

[a] *Center for Quantitative Economics, University of Münster.*

[b] *School of Business and Economics, University of Tübingen.*

Abstract

This online appendix provides comprehensive supplementary material to the main paper in three parts. First, we present an extensive Monte Carlo study that evaluates the pruned skewed Kalman filter and smoother across four carefully designed data-generating processes. The simulation exercises assess state estimation accuracy under different loss functions, analyze computational performance as dimensionality and sample size increase, and examine finite-sample properties of quasi-maximum likelihood estimators for distributional parameters of skewed structural shocks. Our results demonstrate that the pruned skewed Kalman filter achieves numerical accuracy virtually identical to the non-pruned version while delivering speedups of several hundred times, making it computationally feasible for multivariate applications where the non-pruned filter becomes prohibitive. Second, we provide implementation details for Bayesian and maximum likelihood estimation of DSGE models with skew normal shocks, including practical recommendations for mode-finding and posterior sampling. Third, we derive the complete backward recursions for the skewed Kalman smoother, presenting mathematical proofs for the closed skew normal distribution parameters at each smoothing step, with explicit formulas for any time period in the general multivariate case.

Email addresses: `gaygysyz.guljanov@wiwi.uni-muenster.de` (Gaygysyz Guljanov), `willi@mutschler.eu` (Willi Mutschler), `mark.trede@uni-muenster.de` (Mark Trede)

Contents

1. Monte Carlo Study

We conduct a comprehensive Monte Carlo study to evaluate the pruned skewed Kalman filter and smoother across three critical dimensions: state estimation accuracy, computational efficiency, and parameter estimation properties. The study employs four carefully designed state-space models as data-generating processes (DGP), each serving distinct analytical purposes. The first two DGP focus on algorithmic performance assessing how accurately the filter and smoother recover latent states under various loss functions and quantifying the computational burden of likelihood evaluation—as dimensionality and sample size increase. The latter two DGPs examine parameter estimation challenges under different parametrization strategies.

1.1. Data generating processes

DGP(1). The first DGP establishes a baseline univariate system with skewed state transitions:

$$G = 0.8, \quad F = 10, \quad \mu_\varepsilon = 1, \quad \Sigma_\varepsilon = 0.01, \quad \mu_\eta = 0.3, \quad \Sigma_\eta = 0.64, \quad \nu_\eta = 0, \quad \lambda_\eta = -0.89$$

We employ a convenient reparametrization through an auxiliary hyperparameter $\lambda \in (-1, 1)$, expressing the skewness parameters as

$$\Gamma_\eta = \lambda_\eta \Sigma_\eta^{-1/2} \quad \text{and} \quad \Delta_\eta = (1 - \lambda_\eta^2) I_{n_\eta}.$$

This choice elegantly reduces the matrix product $\Delta_\eta + \Gamma'_\eta \Sigma_\eta \Gamma_\eta$ to the identity matrix, yielding closed-form expressions for the unconditional moments of η_t that are also valid in the multivariate case:

$$E[\eta_t] = \mu_\eta + \left(\sqrt{\frac{2}{\pi}} \lambda_\eta {\Sigma_\eta}^{1/2}\right) \mathbf{1}_{n_\eta} \quad \text{and} \quad V[\eta_t] = \Sigma_\eta \left(1 - \frac{2}{\pi} \lambda_\eta^2\right)$$

Arellano-Valle & Azzalini (2008) and Käärik et al. (2015) provide related discussions on the usefulness of this re-parameterization for the skew normal distribution.[1] Beyond computational convenience, this parametrization resolves the fundamental identifiability issue inherent in the CSN distribution and facilitates moment-based estimation strategies (Flecher et al., 2009).

[1] This is without loss of generality. We mainly use this to quickly compute $E[\eta_t]$ and $V[\eta_t]$ as we use these values as input parameters for the *Gaussian Kalman Filter*. In fact, our replication codes contain functions to compute the unconditional mean and the covariance matrix for any valid parameterization of the multivariate CSN distribution. Moreover, in our empirical application we do not use this reparametrization.

DGP(2). The second DGP scales up complexity by introducing a multivariate system with four state and three observable variables with randomly drawn parameters:

$$G = \begin{pmatrix} 0.5488 & 0.1738 & -0.2949 & 0.1534 \\ -0.2864 & 0.1060 & 0.3628 & 0.3334 \\ -0.3898 & -0.0252 & 0.5339 & 0.3163 \\ 0.2389 & 0.1958 & -0.0027 & 0.5519 \end{pmatrix} \qquad F = \begin{pmatrix} -0.7196 & 0.8221 & 0.4602 & -0.6412 \\ -2.0887 & -0.8201 & -1.2380 & 0.3937 \\ 0.6347 & -0.5109 & 0.8476 & 0.6819 \end{pmatrix}$$

$$\Sigma_\varepsilon = \begin{pmatrix} 0.0108 & -0.0276 & -0.0314 \\ -0.0276 & 0.1129 & -0.0025 \\ -0.0314 & -0.0025 & 0.2889 \end{pmatrix} \cdot 10^{-6} \qquad \Sigma_\eta = \begin{pmatrix} 0.0013 & -0.0111 & 0.0116 & -0.0089 \\ -0.0111 & 0.1009 & -0.2301 & 0.1014 \\ 0.0116 & -0.2301 & 3.3198 & -1.0618 \\ -0.0089 & 0.1014 & -1.0618 & 1.0830 \end{pmatrix}$$

$$\mu_\varepsilon = \begin{pmatrix} 0.8565 \\ -0.3010 \\ -0.82705 \end{pmatrix}, \qquad \mu_\eta = \begin{pmatrix} 0.3455 \\ -1.8613 \\ 0.7765 \\ -0.5964 \end{pmatrix}, \qquad \nu_\eta = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \lambda_\eta = 0.89$$

This DGP directly confronts the *curse of increasing skewness dimension* where computational requirements grow exponentially with sample size and the non-pruned filter becomes computationally prohibitive.

DGP(3). The third DGP shifts focus to parameter estimation challenges by abandoning the auxiliary parametrization. It features three observables with measurement error and three states with distinct distributional characteristics on the shock processes:

$$G = \begin{pmatrix} 0.9969 & 0.1256 & -0.4803 \\ -0.8221 & 0.0386 & 0.6687 \\ 0.5605 & 0.6397 & -0.4333 \end{pmatrix}, \quad F = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mu_\varepsilon = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mu_\eta = \begin{pmatrix} 0.3 \\ -0.1 \\ 0.2 \end{pmatrix}, \quad \nu_\eta = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\Sigma_\varepsilon = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot 10^{-4}, \quad \Sigma_\eta = \begin{pmatrix} 0.64 & 0 & 0 \\ 0 & 0.36 & 0 \\ 0 & 0 & 0.49 \end{pmatrix}, \quad \Gamma_\eta = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -6 \end{pmatrix}, \quad \Delta_\eta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The goal is to test the filter's ability to correctly identify and estimate different shocks—one right-skewed, one symmetric (Gaussian), and one left-skewed—with maximum likelihood. This specification is inspired by the multivariate Dynamic Nelson-Siegel term structure model of Diebold et al. (2006), which we estimate in Guljanov et al. (2022). For identification, we impose the standard normalization rule $\nu_\eta = 0$ and $\Delta_\eta = I_{n_\eta}$, then directly estimate μ_η, Σ_η, and Γ_η by maximizing the log-likelihood function.

DGP(4). The fourth DGP operationalizes the DSGE model from the main paper, where we fix the model, stderr and skew parameters to the posterior mean values from Table 2 in the main paper. Structural shocks follow *independent* univariate skew normal distributions which permits an elegant representation via Gupta et al. (2004)'s multivariate formulation: $CSN_{n_\eta,n_\eta}(\mu_\eta, \Sigma_\eta, \Gamma_\eta, 0, I_{n_\eta})$, where Σ_η and Γ_η are diagonal matrices. This offers three key advantages. First, it ensures unique identification of the structural shock distribution which becomes multivariate skew normal. Second, while univariate skew normal distributions constrain individual shock skewness to approximately ± 0.995, the recursive filtering operations generate time-varying skewness parameters for the states and observables that may exceed these bounds, as the CSN framework accommodates more extreme asymmetry through its expanded parameter space. Third, the availability of closed-form univariate moment expressions enables analytical calibration of μ_η to enforce $E[\eta_t] = 0$, thereby preserving the standard zero-mean assumption for structural innovations without numerical optimization. Furthermore, it establishes an analytical one-to-one mapping between interpretable moments (standard deviations and skewness coefficients) and the CSN parameters (diagonal entries of Σ_η and Γ_η), allowing seamless translation between the theoretical parametrization and empirically meaningful quantities. Our Dynare implementation adopts this moment-based parametrization, extending the familiar `stderr` and `corr` syntax with a new `skew` keyword, thereby maintaining continuity with existing DSGE modeling and estimation practices.

1.2. Filter initialization and log-likelihood computation

For DGPs (1) to (3) we set the initial state distribution to a normal one with an initial covariance matrix with 10 on the diagonal, $x_{0|0} \sim CSN(0, 10I_{n_x}, 0, 0, I_{n_x}) = N(0, 10I_{n_x})$, following a suggestion of Harvey & Phillips (1979). For DGP(4) we follow standard practice in DSGE estimation and use the solution to the Lyapunov equation, $\Sigma_{0|0} = G\Sigma_{0|0}G' + Var(\eta_t)$, to compute the initial state distribution's covariance matrix $x_{0|0} \sim CSN(0, \Sigma_{0|0}, 0, 0, I_{n_x}) = N(0, \Sigma_{0|0})$. Our results do not depend on these choices. To compute the log-likelihood function, we make use of the standard predictive decomposition based on the conditional distribution of y_t given y_{t-1}:

$$y_t|y_{t-1} \sim CSN(\hat{y}_{t|t-1}, \Omega_{t|t-1}, K_{t-1}^{Skewed}, \nu_{t|t-1}, \Delta_{t|t-1} + (\Gamma_{t|t-1} - K_{t-1}^{Skewed}F)\Sigma_{t|t-1}\Gamma_{t|t-1}')$$

where $\hat{y}_{t|t-1} = F\mu_{t|t-1} + \mu_\epsilon$ is the predicted value and $\Omega_{t|t-1} = F\Sigma_{t|t-1}F' + \Sigma_\epsilon$ is the prediction-error covariance matrix of the Gaussian Kalman filter. Evidently, for $\Gamma_{t|t-1} = 0$ the CSN log-likelihood collapses to the Gaussian log-likelihood.

1.3. Filtering and smoothing performance

For forecasting, it is helpful to condense the filtered distribution $x_{t|t}$ (and similarly the smoothed distribution $x_{t|T}$) into a point estimator. Since the CSN distribution is asymmetric, the expectation $E[x_{t|t}]$ is one potential, but not necessarily the best point estimator. Let $L[\tilde{x}_t, x_t]$ denote the loss function for a point estimator $\tilde{x}_t$ if the true value is x_t. Depending on the loss function, different point estimators will minimize the expected loss. Of course, if the loss function is quadratic, i.e. $L_2[\tilde{x}_t, x_t] = (\tilde{x}_t - x_t)^2$, the expected loss is minimal if $\tilde{x}_t = E[x_{t|t}]$. If the loss function is $L_1[\tilde{x}_t, x_t] = |\tilde{x}_t - x_t|$, the best point estimator is the median of $x_{t|t}$. And the asymmetric loss function

$$L_q[\tilde{x}_t, x_t] = \begin{cases} a|\tilde{x}_t - x_t| & \text{if } x_t > \tilde{x}_t \\ b|\tilde{x}_t - x_t| & \text{if } x_t \leq \tilde{x}_t \end{cases}$$

results in the $a/(a+b)$-quantile of $x_{t|t}$ as point estimator.

We start by simulating $R = 2400$ sample paths for x_t and y_t of different length $T = \{40, 80, 110\}$ (plus a burn-in of 100 periods). The shocks η_t are drawn from the CSN distribution, whereas the measurement errors ε_t are drawn from the normal distribution. We compute the expected losses to assess how well the *Pruned Skewed Kalman Filter and Smoother* estimate the unobserved state variables in comparison to the conventional *Skewed* or *Gaussian Kalman Filters and Smoothers*. That is, for each sample $r = 1, \ldots, R$ the loss is computed as

$$\text{Loss}^{(r)} := \sum_{t=20}^{T} L[\tilde{x}_t^{(r)}, x_t^{(r)}]$$

where in the univariate case L is any of the three loss functions L_1, L_2 and L_q (with $a = 1$ and $b = 4$) under consideration, while in the multivariate case we focus only on L_2 as multivariate versions of L_1 and L_q are not readily available. Note that in order to avoid too large an impact of the initial distribution $x_{0|0}$, the losses are calculated after a burn-in phase of 20 periods. The expected loss is then estimated by averaging over all replications

$$\text{Expected Loss} = \frac{1}{R}\sum_{r=1}^{R} \text{Loss}^{(r)}.$$

Tables 1 and 2 report the *Expected Loss* and the 5th and 95th percentiles of *Loss*$^{(r)}$ of our Monte-Carlo simulation exercise for the different variants of both filters and smoothers. Several things are worth pointing out.

First, the *Skewed Kalman Filter and Smoother* are superior to the *Gaussian Kalman Filter and Smoother* in all cases. Even though the better performance is rather small in the univariate case, it becomes really measurable in the multivariate case. This is not surprising, since the closed skew normal distribution deviates

Table 1: Expected losses for filtered states

DGP	T	LOSS	GAUSSIAN	NO PRUNING	TOL=1e-6	TOL=1e-4	TOL=1e-2
(1)	40	L_1	0.166535064 [0.1236;0.2142]	0.166527075 [0.1235;0.2144]	0.166527075 [0.1235;0.2144]	0.166527075 [0.1235;0.2144]	0.166527075 [0.1235;0.2144]
(1)	40	L_2	0.002080850 [0.0012;0.0033]	0.002080707 [0.0012;0.0033]	0.002080707 [0.0012;0.0033]	0.002080707 [0.0012;0.0033]	0.002080707 [0.0012;0.0033]
(1)	40	L_q	0.293173611 [0.2168;0.3852]	0.293160294 [0.2167;0.3852]	0.293160294 [0.2167;0.3852]	0.293160294 [0.2167;0.3852]	0.293160288 [0.2167;0.3852]
(1)	80	L_1	0.486149061 [0.4151;0.5639]	0.486122323 [0.4151;0.5638]	0.486122323 [0.4151;0.5638]	0.486122323 [0.4151;0.5638]	0.486122327 [0.4151;0.5638]
(1)	80	L_2	0.006087964 [0.0045;0.0080]	0.006087485 [0.0045;0.0080]	0.006087485 [0.0045;0.0080]	0.006087485 [0.0045;0.0080]	0.006087485 [0.0045;0.0080]
(1)	80	L_q	0.853449643 [0.7129;1.0017]	0.853414297 [0.7127;1.0008]	0.853414297 [0.7127;1.0008]	0.853414297 [0.7127;1.0008]	0.853414287 [0.7127;1.0008]
(1)	110	L_1	0.724620237 [0.6368;0.8186]	0.724598686 [0.6367;0.8185]	0.724598686 [0.6367;0.8185]	0.724598686 [0.6367;0.8185]	0.724598685 [0.6367;0.8185]
(1)	110	L_2	0.009073018 [0.0071;0.0113]	0.009072577 [0.0071;0.0113]	0.009072577 [0.0071;0.0113]	0.009072577 [0.0071;0.0113]	0.009072577 [0.0071;0.0113]
(1)	110	L_q	1.272405929 [1.1026;1.4533]	1.272363081 [1.1024;1.4536]	1.272363081 [1.1024;1.4536]	1.272363081 [1.1024;1.4536]	1.272363058 [1.1024;1.4536]
(2)	40	L_2	4.23932054 [2.1343;6.9488]	4.17299006 [2.1172;6.9381]	4.17299000 [2.1172;6.9381]	4.17298994 [2.1173;6.9382]	4.17450665 [2.0989;6.9805]
(2)	80	L_2	12.30937668 [8.4001;17.0700]	-	12.11085307 [8.3181;16.9039]	12.11085498 [8.3181;16.9048]	12.11665912 [8.3003;16.9054]
(2)	110	L_2	18.39547677 [13.4673;24.0186]	-	18.10271658 [13.1829;23.6744]	18.10272323 [13.1834;23.6743]	18.11208988 [13.2441;23.6814]

Note: Lower is better. 5th and 95th percentiles in square brackets.

only mildly from symmetry and normality and the conventional Kalman filter and smoother are naturally optimal in its domain, i.e. when data is very close to normal. Nevertheless, in the more general case, the conventional Kalman filter and smoother simply neglect the skewed behavior; while the *Skewed Kalman Filter* embeds normality as a special case.

Second, our pruning algorithm is very accurate and numerically almost equivalent to the non-pruned *Skewed Kalman Filter* (up to the twelfth digit in the univariate case and up to the 5th digit in the multivariate case). We also calculate the Kullback-Leibler divergence measures (KL) to quantify the expected loss of information due to pruning (relative to the SKF without pruning). For pruning tolerance *tol* we define

$$KL(tol) = E\left[\log \frac{f(x_1, \ldots, x_T|SKF)}{f(x_1, \ldots, x_T|PSKF(tol))}\right]$$

where $f(x_1, \ldots, x_T|SKF)$ is the likelihood of the states without pruning while $f(x_1, \ldots, x_T|PSKF(tol))$ is the likelihood with pruning tolerance tol. The expectation is approximated by averaging over all samples $r = 1, \ldots, R$. Since computing the skewed Kalman filter without pruning quickly becomes prohibitive in the multivariate setting when the observation window grows, we compute KL only for the univariate DGP (1). For sample sizes 40, 80 and 110, and pruning thresholds of 10^{-2}, 10^{-4} and 10^{-6}, the value of $KL(tol)$ never

Table 2: Expected losses for smoothed states

DGP	T	LOSS	GAUSSIAN	NO PRUNING	TOL=1e-6	TOL=1e-4	TOL=1e-2
(1)	40	L_1	0.166539615 [0.1239;0.2143]	0.166528591 [0.1236;0.2145]	0.166528591 [0.1236;0.2145]	0.166528596 [0.1236;0.2145]	0.166528596 [0.1236;0.2145]
(1)	40	L_2	0.002080868 [0.0012;0.0033]	0.002080555 [0.0012;0.0033]	0.002080555 [0.0012;0.0033]	0.002080555 [0.0012;0.0033]	0.002080555 [0.0012;0.0033]
(1)	40	L_q	0.293186237 [0.2162;0.3848]	0.293159674 [0.2163;0.3847]	0.293159674 [0.2163;0.3847]	0.293159669 [0.2163;0.3847]	0.293159669 [0.2163;0.3847]
(1)	80	L_1	0.486116200 [0.4153;0.5645]	0.486083757 [0.4156;0.5644]	0.486083757 [0.4156;0.5644]	0.486083761 [0.4156;0.5644]	0.486083761 [0.4156;0.5644]
(1)	80	L_2	0.006087517 [0.0045;0.0080]	0.006086645 [0.0045;0.0080]	0.006086645 [0.0045;0.0080]	0.006086645 [0.0045;0.0080]	0.006086645 [0.0045;0.0080]
(1)	80	L_q	0.853428859 [0.7133;1.0015]	0.853347301 [0.7125;1.0010]	0.853347301 [0.7125;1.0010]	0.853347286 [0.7125;1.0010]	0.853347286 [0.7125;1.0010]
(1)	110	L_1	0.724563642 [0.6363;0.8186]	0.724528917 [0.6357;0.8184]	0.724528917 [0.6357;0.8184]	0.724528915 [0.6357;0.8184]	0.724528915 [0.6357;0.8184]
(1)	110	L_2	0.009072162 [0.0070;0.0113]	0.009071158 [0.0070;0.0113]	0.009071158 [0.0070;0.0113]	0.009071158 [0.0070;0.0113]	0.009071158 [0.0070;0.0113]
(1)	110	L_q	1.272365795 [1.1024;1.4536]	1.272252412 [1.1017;1.4534]	1.272252412 [1.1017;1.4534]	1.272252388 [1.1017;1.4534]	1.272252388 [1.1017;1.4534]
(2)	40	L_2	0.37817718 [0.1077;1.0473]	0.37728799 [0.1100;1.0227]	0.37728800 [0.1100;1.0227]	0.37728833 [0.1100;1.0226]	0.37740658 [0.1099;1.0301]
(2)	80	L_2	0.66853761 [0.3625;1.3612]	-	0.66267846 [0.3611;1.3307]	0.66267825 [0.3611;1.3306]	0.66275740 [0.3612;1.3280]
(2)	110	L_2	0.88605162 [0.5568;1.5668]	-	0.87956632 [0.5538;1.5548]	0.87956592 [0.5538;1.5549]	0.87960797 [0.5550;1.5550]

Note: Lower is better. 5th and 95th percentiles in square brackets.

exceeds 2.4×10^{-5}, indicating that there is hardly any loss of information by pruning.

Third, the pruning threshold does not matter measurably in the univariate case and makes only a small numerical difference in multivariate settings. Clearly, the closer the tolerance is to 0, i.e. to the non-pruned filter and smoother, the more accurate we estimate the states. However, as we have argued above the non-pruned version of the filter and smoother is only feasible in the univariate case, while in multivariate settings we manage to deal with the numerical challenges for very small sample sizes only. Our pruning algorithm, on the other hand, is able to overcome this problem. Even with very low tolerance thresholds we are able to compute the filtering and smoothing steps without running into the curse of increasing skewness dimensionality. We conclude that overall both the *Pruned Skewed Kalman Filter and Smoother* perform well in terms of accuracy.

Fourth, the *smoother* consistently outperforms the *filter* across all specifications, as expected from theory since smoothing uses both past and future information. The improvement is modest but consistent: for instance, in DGP(2) at T=40, the smoothed L_2-loss is more than 90% lower than the filtered loss.

Fifth, the relative performance gains of the skewed approach over the Gaussian are remarkably stable across different sample sizes, suggesting that the benefits persist regardless of the observation window length.

Sixth, the confidence bands (5th and 95th percentiles) remain relatively tight around the expected losses, indicating robust and consistent performance across Monte Carlo replications—this stability is particularly notable for the pruned versions.

Finally, the asymmetric loss function L_q with parameters $a = 1$ and $b = 4$ yields consistently higher losses than L_1 and L_2, reflecting its heavier penalization of underestimation relative to overestimation—yet even under this more demanding criterion, the skewed filters maintain their advantage over Gaussian alternatives.

However, there is a trade-off between accuracy and speed, which we analyze next.

1.4. Computational performance

We benchmark the *Pruned Skewed Kalman Filter* (PSKF) by wall-clock time. Table 3 reports the time in **seconds** required to compute **1000 evaluations** of the log-likelihood function of univariate DGP (1) and multivariate DGP (2) for different sample sizes. Across all methods, computation time naturally rises with $T \in \{50, 100, 150, 200, 250\}$. The non-pruned *Skewed Kalman Filter* (SKF) illustrates the curse of dimensionality most starkly. In the univariate DGP, its time jumps from 26 seconds at $T = 50$ to 5408 seconds at $T = 250$ (factor of about 208); in the multivariate DGP, it is already extremely slow at $T = 100$ (10903 seconds) and becomes infeasible for $T \geq 150$. By contrast, the PSKF scales similarly to the Gaussian Kalman filter (KF): in DGP (1) the KF's runtime rises by a factor of 3.93 from $T = 50$ to $T = 250$, while the PSKF's runtime rises by 4.16 with $tol = 10^{-6}$ and 4.25 with $tol = 10^{-2}$. In DGP (2) the corresponding growth factors are 6.17 (KF), 5.38 (PSKF, 10^{-6}), and 4.95 (PSKF, 10^{-2}).

Where the non-pruned SKF becomes impractical (or impossible) beyond small T in DGP (2), the PSKF stays tractable. It delivers massive speedups relative to the *non-pruned* SKF even in the feasible cases: in DGP (1) at $T = 250$, $tol = 10^{-2}$ is roughly 340 times faster; in DGP (2) at $T = 100$, it is about 300 to 640 times faster depending on tolerance. Obviously, looser pruning thresholds are faster. In DGP (1), $tol = 10^{-2}$ is about 1.5 times faster than 10^{-6} uniformly across T (e.g. 15.84 vs. 23.84 seconds at $T = 250$). In DGP (2), the speedup is about 2.0 to 2.2 times (e.g. 41.45 vs. 90.99 seconds at $T = 250$). Tightening from 10^{-4} to 10^{-6} imposes a small penalty in the univariate case (only $2 - 6\%$ extra time over all T), but a sizable one in the multivariate case ($32 - 63\%$ extra time). These patterns match the accuracy-speed trade-off documented in the previous section.

Naturally, the Gaussian Kalman Filter remains the speed champion, but the absolute overhead of the PSKF is modest in per-evaluation terms. Because the table reports 1000 evaluations, at $T = 250$ the DGP (1) KF costs approximately 1.0 millisecond per evaluation, while the PSKF costs approximately 15.8 milliseconds with $tol = 10^{-2}$ and 23.8 milliseconds with $tol = 10^{-6}$. In DGP (2), the gap is even narrower in relative terms: 5.2 milliseconds (KF) versus 41.5 milliseconds (PSKF, 10^{-2}). The reported 5th-95th percentile bands

Table 3: Computation time

DGP	T	GAUSSIAN	NO PRUNING	TOL=1e-6	TOL=1e-4	TOL=1e-2
(1)	50	0.2555 [0.18;0.39]	26.4004 [19.85;42.68]	5.7352 [4.26;8.55]	5.4255 [4.15;8.35]	3.7236 [2.82;5.74]
(1)	100	0.4236 [0.34;0.63]	178.4764 [164.04;272.46]	9.3932 [8.56;14.91]	9.2007 [8.30;14.70]	6.2735 [5.61;10.14]
(1)	150	0.5988 [0.51;0.92]	764.5933 [730.60;860.85]	13.8686 [13.08;14.44]	13.5295 [12.72;14.12]	9.2433 [8.61;9.72]
(1)	200	0.7869 [0.68;1.28]	2276.7823 [2208.36;2374.31]	18.7564 [17.65;19.54]	18.3690 [17.25;19.01]	12.5310 [11.65;13.20]
(1)	250	1.0029 [0.87;1.65]	5407.6977 [5292.97;5522.95]	23.8373 [22.29;24.89]	23.4432 [21.77;24.56]	15.8381 [14.71;16.57]
(2)	50	0.8439 [0.71;1.46]	554.8769 [518.17;660.16]	16.9193 [15.82;17.94]	12.8243 [11.90;15.25]	8.3821 [7.74;9.32]
(2)	100	1.6507 [1.42;2.99]	10902.8252 [9388.04;13659.95]	36.6296 [32.88;42.71]	26.8729 [24.38;30.29]	17.0615 [15.77;18.32]
(2)	150	3.3166 [2.65;5.69]	-	56.8434 [49.07;89.69]	35.2654 [30.71;55.59]	26.1939 [22.86;41.05]
(2)	200	4.6808 [3.50;7.50]	-	80.1512 [65.44;118.95]	49.6175 [40.84;74.17]	36.6851 [30.44;54.95]
(2)	250	5.2042 [4.35;9.10]	-	90.9946 [81.35;143.74]	55.7294 [50.84;88.30]	41.4539 [37.92;64.54]

Note: Time in seconds to compute 1000 evaluations of the log-likelihood function on AMD EPYC 7402P (24 cores, 96 GB RAM). 5th and 95th percentiles in square brackets.

for the PSKF are modest in absolute terms (typically a few seconds) relative to the orders-of-magnitude cost differences across methods, indicating stable run-to-run behavior.

To sum up, the PSKF is roughly 8 to 24 times slower than the KF depending on dimension and tolerance—well within the realm of standard likelihood-based estimation. Other approaches to evaluate the likelihood, such as Sequential Monte Carlo, are typically much slower by a factor of several hundred or thousand. Combining these results with the accuracy study, we recommend $tol \in [10^{-2}, 10^{-4}]$ as a pragmatic default for multivariate models: 1% when speed is paramount and 0.01% when extra accuracy is desired. In univariate models, tightening to 10^{-6} is feasible at essentially no additional cost relative to 10^{-4} and may be preferred if one wants the most conservative pruning.

In the next section, we examine the finite-sample properties of quasi-maximum-likelihood estimators for the CSN shock parameters. Before turning there, a few implementation remarks are in order because they directly affect the reported computation times. Our implementations of both the *Gaussian Kalman Filter* as well as the *Pruned Skewed Kalman Filter* are deliberately textbook-style (specifically in the Monte-Carlo study) to keep the algorithmic structure transparent and comparable. Substantial speedups are possible—for example, avoid explicit matrix inverses, apply Chandrasekhar recursions, or use steady-state filtering since K^{Gauss} converges rather quickly and K^{Skew} also converges, albeit somewhat slower. In experiments with these modifications, we reduced computing times by at least one half. To keep comparisons clean, however, all results reported here use non-optimized, textbook implementations; the *Gaussian Kalman Filter* could likewise be accelerated under the same optimizations.

1.5. Estimation performance

Our final simulation exercise examines finite-sample estimation performance by generating $R = 1200$ datasets from the multivariate DGPs (3) and (4). We focus exclusively on estimating the distributional parameters of η_t—specifically μ_η, $\log(diag(\Sigma_\eta))$, and $diag(\Gamma_\eta)$ for DGP (3), and the standard deviation and skewness coefficients for DGP (4)—while keeping all other parameters fixed at their true values.[2] Following Atkinson et al. (2019), we assess estimation accuracy through multiple metrics: the average estimate, the 5th and 95th percentiles, and the normalized root-mean square-error (NRMSE). The NRMSE for parameter j and Kalman filter variant f is defined as:

$$NRMSE_f^j = \frac{1}{\theta_j}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{j,f,r} - \theta_j\right)^2}$$

where $\hat{\theta}_{j,f,r}$ denotes the corresponding estimate for dataset r and θ_j the true parameter value. This normalization facilitates comparison across parameters with different scales.

Table **??** presents results on parameter estimates across three Kalman filter variants and sample sizes, with each cell reporting the average estimate, the [5th, 95th] percentiles, and the {NRMSE}. The results demonstrate three key findings about the *Pruned Skewed Kalman Filter*'s ability to estimate shock parameters.

First, the method exhibits remarkable robustness: estimates remain virtually identical across pruning thresholds (10^{-6} vs 10^{-2}), and the parameter distributions consistently center around true values with narrowing confidence intervals as sample size increases. In DGP (3), the filter correctly identifies both the positive skewness of $\eta_{1,t}$, the negative skewness of $\eta_{3,t}$, and crucially, the Gaussianity of $\eta_{2,t}$—confirming that the method does not introduce spurious skewness.

Second, the NRMSE analysis reveals a clear hierarchy in estimation precision: scale parameters Σ_η are most accurately estimated ($NRMSE \in [0.10, 0.20]$), followed by location parameters μ_η ($NRMSE \in [0.20, 0.40]$ for skewed components), while skewness parameters Γ_η prove most challenging ($NRMSE \in [0.40, 0.80]$). This pattern reflects the inherent difficulty of estimating higher-order moments, though accuracy improves substantially with sample size—for instance, the NRMSE for $\Gamma_{\eta,11}$ falls from 0.811 at $T = 100$ to 0.395 at $T = 200$.

Third, while the Gaussian Kalman filter seems to severely mis-estimate the underlying distributional parameters μ_η and Σ_η in DGP (3), these "biased" estimates actually correspond to consistent estimates

[2] We log-transform Σ_η to avoid non-negativity constraints during estimation. The reported estimates are re-transformed.

Table 4: Distribution of parameter estimates DGP (3)

Param	Truth	*Pruned Skewed KF (1e-6)*			*Pruned Skewed KF (1e-2)*			*Gaussian KF*		
		100	150	200	100	150	200	100	150	200
$[\mu_\eta]_1$	0.30	0.302 [0.19;0.43] {0.261}	0.299 [0.21;0.40] {0.193}	0.296 [0.22;0.38] {0.173}	0.302 [0.19;0.43] {0.260}	0.299 [0.21;0.40] {0.193}	0.297 [0.22;0.39] {0.173}	0.922 [0.83;1.01] {2.081}	0.924 [0.86;1.00] {2.083}	0.921 [0.86;0.98] {2.075}
$[\mu_\eta]_2$	-0.10	−0.101 [−0.20;−0.00] {−0.610}	−0.099 [−0.18;−0.02] {−0.480}	−0.100 [−0.17;−0.03] {−0.415}	−0.101 [−0.20;−0.00] {−0.615}	−0.099 [−0.18;−0.02] {−0.484}	−0.100 [−0.17;−0.03] {−0.418}	−0.101 [−0.20;−0.00] {−0.615}	−0.099 [−0.18;−0.02] {−0.484}	−0.100 [−0.17;−0.03] {−0.418}
$[\mu_\eta]_3$	0.20	0.195 [0.08;0.29] {0.321}	0.200 [0.11;0.27] {0.245}	0.198 [0.13;0.26] {0.214}	0.195 [0.08;0.29] {0.321}	0.200 [0.11;0.27] {0.245}	0.198 [0.13;0.26] {0.214}	−0.345 [−0.42;−0.28] {2.734}	−0.343 [−0.40;−0.28] {2.723}	−0.344 [−0.39;−0.29] {2.723}
$[\Sigma_\eta]_{11}$	0.64	0.654 [0.44;0.87] {0.203}	0.658 [0.50;0.84] {0.168}	0.656 [0.51;0.81] {0.148}	0.654 [0.44;0.87] {0.203}	0.658 [0.50;0.84] {0.168}	0.656 [0.51;0.81] {0.148}	0.260 [0.19;0.34] {0.597}	0.262 [0.21;0.32] {0.593}	0.261 [0.21;0.31] {0.594}
$[\Sigma_\eta]_{22}$	0.36	0.351 [0.27;0.44] {0.141}	0.353 [0.29;0.42] {0.117}	0.354 [0.30;0.41] {0.101}	0.351 [0.27;0.44] {0.141}	0.353 [0.29;0.42] {0.117}	0.354 [0.30;0.41] {0.101}	0.351 [0.27;0.44] {0.141}	0.353 [0.29;0.42] {0.117}	0.354 [0.30;0.41] {0.101}
$[\Sigma_\eta]_{33}$	0.49	0.491 [0.33;0.64] {0.197}	0.494 [0.37;0.62] {0.160}	0.489 [0.39;0.60] {0.133}	0.491 [0.33;0.64] {0.197}	0.494 [0.37;0.62] {0.160}	0.489 [0.39;0.60] {0.133}	0.193 [0.14;0.25] {0.609}	0.195 [0.15;0.24] {0.605}	0.193 [0.16;0.23] {0.607}
$[\Gamma_\eta]_{11}$	5.00	6.178 [3.31;11.47] {0.811}	5.712 [3.56;9.22] {0.513}	5.631 [3.78;8.63] {0.395}	6.173 [3.31;11.43] {0.804}	5.712 [3.56;9.21] {0.513}	5.630 [3.78;8.62] {0.394}			
$[\Gamma_\eta]_{22}$	0.00	−0.001 [−0.01;0.00] { }	−0.001 [−0.01;0.00] { }	−0.000 [−0.00;0.00] { }	−0.001 [−0.01;0.00] { }	−0.001 [−0.01;0.00] { }	−0.001 [−0.01;0.00] { }			
$[\Gamma_\eta]_{33}$	-6.00	−7.538 [−15.35;−3.76] {−0.741}	−6.849 [−11.09;−4.21] {−0.476}	−6.581 [−9.82;−4.30] {−0.376}	−7.535 [−15.36;−3.77] {−0.739}	−6.848 [−11.08;−4.21] {−0.474}	−6.581 [−9.82;−4.30] {−0.377}			

Note: Cells contain average on top, [5,95] percentiles in square brackets, and {NRMSE} in curly brackets. For $[\Gamma_\eta]_{22}$, NRMSE is not defined due to division by zero.

of the unconditional moments $E[\eta_t]$ and $V[\eta_t]$. In our setup, these equal $[0.9192; -0.1000; -0.3433]$ and $diag([0.2565; 0.3600; 0.1948])$, respectively, which align roughly with the Gaussian filter's estimates for μ_η and Σ_η. Thus, the Gaussian filter remains useful when only first and second moments are of interest, though it fundamentally cannot capture the distributional characteristics that drive asymmetric dynamics. The *Pruned Skewed Kalman Filter*, by contrast, provides a more complete characterization by nesting Gaussianity as a special case while accurately recovering skewed distributions when present.

Table 4 extends the analysis to the DSGE-based DGP (4). The standard deviation parameters are estimated with high precision across all four shocks. For the skewness parameters, the filter successfully recovers the negative skewness of η_a and η_z as well as the positive skewness of η_r, with estimates centering close to their true values and confidence intervals narrowing substantially as sample size increases.

2. Computational remarks for estimating Ireland (2004) with skew normal shocks

All estimations and simulations were performed with Dynare 7.0 and MATLAB R2025b on an Apple MacBook Pro equipped with an M2 Max chip and 64 GB RAM.

2.1. Filter initialization and likelihood penalization

Based on our Monte Carlo evidence, we prune skewness dimensions below a 1% threshold. The initial distribution for the prediction-error decomposition of the likelihood is set to a normal distribution with mean zero and initial forecast-error covariance equal to the unconditional variance of the state variables (solution to the Lyapunov equation). We penalize the likelihood (and posterior) function in several scenarios, such as when the Blanchard & Kahn (1980) conditions are violated (i.e. a DSGE specific generalization of eigenvalues of G being outside the unit circle), the covariance matrix of η_t is not positive semi-definite, or the skewness parameters exceed theoretical limits.

2.2. Grid search and maximum likelihood optimization

To find the mode of the likelihood, we use a sophisticated search for initial parameter values—a critical step in any maximum likelihood estimation exercise. We start with values from Ireland (2004) for model and standard error parameters, then construct an evenly spaced grid of skewness parameters for all four shocks. For each value on the grid, we compute the log-likelihood, while holding model parameters and standard errors of shocks fixed at their Gaussian estimates. By evaluating over 50,000 combinations (spanning various skewness scenarios between the shocks), we identify the best five parameter sets. These serve as starting points for further numerical optimization of both standard errors and skewness parameters, with model parameters held fixed. The resulting shock parameter estimates are merged with the Gaussian model parameter estimates to form our final initial values for the actual estimation. Equipped with these, we minimize the negative log-likelihood function over all parameters.

2.3. Posterior mode finding and MCMC sampling

For Bayesian estimation, a similar approach finds the posterior mode, though the Hessian at the mode can be non-positive definite.[3] To address this, we propose two solutions: (i) Run a Monte Carlo-based optimization

[3]The Metropolis-Hastings algorithm does not require starting exactly at the posterior mode; it only needs a starting point with high posterior density and an estimate of the proposal distribution's covariance matrix. Starting at the the mode is, however, beneficial for the acceptance rate and convergence speed and has become standard practice in the Bayesian estimation of DSGE models literature. Likewise, standard practice is to present Bayesian results using the RWMH algorithm; however, we would like to mention that the posterior distribution obtained by RWMH with 2,000,000 draws closely matches the one obtained via a Slice sampler with 40,000 draws. This is noteworthy, because the Slice sampler typically yields Markov chains with lower autocorrelation than RWMH (so less draws required, albeit each draw requires more function evaluations), but more importantly, it avoids both the time-consuming and often frustrating mode-finding step as well as additional fine-tuning to achieve a specific acceptance rate. Results are accessible in the replication package.

routine to locate a high-density region for initializing the Metropolis-Hastings algorithm and to estimate the posterior covariance matrix. This method is, however, very time-consuming. (ii) Use a (short) Slice sampler to estimate the mode and posterior covariance matrix directly, as it requires no fine-tuning and is very robust in terms of dealing with multi-modality and high-dimensional parameter spaces. Specifically, we run multiple short Slice sampler chains in parallel (e.g., 8 chains with 250 draws each or less), then use the combined draws to determine the mode and covariance matrix of the posterior distribution which then serves for the initialization of the RWMH algorithm. Our code demonstrates that both approaches yield very similar initialization matrices; therefore, we strongly recommend the Slice sampler method for it is faster and more general applicable—not just for the PSKF but for Bayesian estimation in general. Subsequently, we generate 2,000,000 draws across 8 parallel chains with the RWMH algorithm, allocating half of the samples for burn-in, and fine-tuning the proposal distribution to achieve an acceptance rate of around 30% for each chain.

2.4. Estimation with free policy parameters

Lastly, we also estimated model variants with free α_x and α_π parameters; in those cases, we apply a logit transformation to place them on an unbounded domain and—following Ireland (2004)—use one-sided finite differences of the inverse Hessian to compute maximum likelihood standard errors. Bayesian posterior mode finding becomes somewhat more challenging and heavily time-intensive in that setup due to multi-modality.

3. Skewed Kalman Smoother

In this section we derive the smoothing step of the skewed Kalman filter in general.

3.1. Introduction and notations

Derivation of the smoothing step is a tedious task, also in terms of notations. In order to carry out this task as neatly as possible, we will proceed as follows. We first set notations and abbreviations. Afterwards, in a separate section, we show some useful derivations which we use later on. Then the derivation of the smoothing step proceeds with finding smoothed variables of periods $T-1$, $T-2$, $T-3$ and $T-4$, recursively. From what we learn by deriving the above four periods manually, we devise general formulas for any time period.

Let us start with with notations and abbreviations. Define

$$\begin{aligned} J_t &= \Sigma_{t|t} G' \Sigma_{t+1|t}^{-1} \\ K_t &= \Sigma_{t+1|t} F' (F \Sigma_{t+1|t} F' + \Sigma_\varepsilon)^{-1} (y_{t+1} - F \mu_{t+1|t}). \end{aligned}$$

If the transition equation matrix G does not have full rank, the inverse $\Sigma_{t+1|t}^{-1}$ does not exist. It can, however, be replaced by the pseudo-inverse. The same is true for the inverse matrices in the derivations below. In the notation, we do not distinguish between the inverse and the pseudo-inverse matrices. We do, however, assume that Σ_ε has full rank.

For any t, we know the prediction step distribution (see Rezaie & Eidsvik (2014))

$$x_{t+1|t} \sim CSN_{p, q_t + q_\eta}(\mu_{t+1|t}, \Sigma_{t+1|t}, \Gamma_{t+1|t}, \nu_{t+1|t}, \Delta_{t+1|t})$$

where

$$\begin{aligned} \mu_{t+1|t} &= G \mu_{t|t} + \mu_\eta \\ \Sigma_{t+1|t} &= G \Sigma_{t|t} G' + \Sigma_\eta \\ \Gamma_{t+1|t} &= \begin{pmatrix} \Gamma_{t|t} J_t \\ \Gamma_\eta \Sigma_\eta \Sigma_{t+1|t}^{-1} \end{pmatrix} \\ \nu_{t+1|t} &= \begin{pmatrix} \nu_{t|t} \\ \nu_\eta \end{pmatrix} \\ \Delta_{t+1|t} &= \begin{pmatrix} \Delta_{t|t} + \Gamma_{t|t} \Sigma_{t|t} \Gamma_{t|t}' - \Gamma_{t|t} J_t G \Sigma_{t|t} \Gamma_{t|t}' & \Gamma_{t|t} J_t \Sigma_\eta \Gamma_\eta' \\ \Gamma_\eta \Sigma_\eta J_t' \Gamma_{t|t}' & \Delta_\eta + \Gamma_\eta \Sigma_\eta \Gamma_\eta' - \Gamma_\eta \Sigma_\eta \Sigma_{t+1|t}^{-1} \Sigma_\eta \Gamma_\eta' \end{pmatrix} \end{aligned}$$

and the update step distribution (see Rezaie & Eidsvik (2014))

$$x_{t+1|t+1} \sim CSN_{p,q_t}(\mu_{t+1|t+1}, \Sigma_{t+1|t+1}, \Gamma_{t+1|t+1}, \nu_{t+1|t+1}, \Delta_{t+1|t+1})$$

where

$$\begin{aligned}
\mu_{t+1|t+1} &= \mu_{t+1|t} + K_t \\
\Sigma_{t+1|t+1} &= \Sigma_{t+1|t} - \Sigma_{t+1|t}F'(F\Sigma_{t+1|t}F' + \Sigma_\varepsilon)^{-1}F\Sigma_{t+1|t} \\
\Gamma_{t+1|t+1} &= \Gamma_{t+1|t} \\
\nu_{t+1|t+1} &= \nu_{t+1|t} - \Gamma_{t+1|t}K_t \\
\Delta_{t+1|t+1} &= \Delta_{t+1|t}.
\end{aligned}$$

3.2. Preliminary theorems and lemmas

Let $\mathcal{F}_t$ denote the information set at time t.

Theorem 1.

$$x_t|x_{t+1}, \mathcal{F}_t \sim CSN(\mu_{dt}, \Sigma_{dt}, \Gamma_{dt}, \nu_{dt}, \Delta_{dt})$$

with

$$\begin{aligned}
\mu_{dt} &= \mu_{t|t} + \Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}(x_{t+1} - G\mu_{t|t} - \mu_\eta) \\
\Sigma_{dt} &= \Sigma_{t|t} - \Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}G\Sigma_{t|t} \\
\Gamma_{dt} &= \begin{pmatrix} \Gamma_{t|t} \\ -\Gamma_\eta G \end{pmatrix} \\
\nu_{dt} &= \begin{pmatrix} \nu_{t|t} \\ \nu_\eta \end{pmatrix} - \begin{pmatrix} \Gamma_{t|t}\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1} \\ \Gamma_\eta - \Gamma_\eta G\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1} \end{pmatrix}(x_{t+1} - G\mu_{t|t} - \mu_\eta) \\
&= \begin{pmatrix} \nu_{t|t} - \Gamma_{t|t}\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}(x_{t+1} - G\mu_{t|t} - \mu_\eta) \\ \nu_\eta - \Gamma_\eta + \Gamma_\eta G\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}(x_{t+1} - G\mu_{t|t} - \mu_\eta) \end{pmatrix} \\
\Delta_{dt} &= \Delta_c \\
&= \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{pmatrix}.
\end{aligned}$$

Proof. Our starting point is the equation

$$\begin{pmatrix} x_{t|t} \\ x_{t+1|t} \end{pmatrix} = \begin{pmatrix} I \\ G \end{pmatrix} x_{t|t} + \begin{pmatrix} 0 \\ \eta_t \end{pmatrix}.$$

which is equivalent to

$$\begin{pmatrix} x_{t|t} \\ x_{t+1|t} \end{pmatrix} = \underbrace{\begin{pmatrix} I & 0 \\ G & I \end{pmatrix}}_{\equiv A} \begin{pmatrix} x_{t|t} \\ \eta_t \end{pmatrix}.$$

The joint distribution is (we drop the time index)

$$\begin{pmatrix} x_{t|t} \\ \eta_t \end{pmatrix} \sim \mathrm{CSN}(\mu_j,\ \Sigma_j,\ \Gamma_j,\ \nu_j,\ \Delta_j)$$

where

$$\begin{aligned}
\mu_j &= \begin{bmatrix} \mu_{t|t} \\ \mu_\eta \end{bmatrix} \\
\Sigma_j &= \begin{bmatrix} \Sigma_{t|t} & 0 \\ 0 & \Sigma_\eta \end{bmatrix} \\
\Gamma_j &= \begin{bmatrix} \Gamma_{t|t} & 0 \\ 0 & \Gamma_\eta \end{bmatrix} \\
\nu_j &= \begin{bmatrix} \nu_{t|t} \\ \nu_\eta \end{bmatrix} \\
\Delta_j &= \begin{bmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{bmatrix}
\end{aligned}$$

Then, take the linear transformation of the above joint distribution with matrix A.

$$A \begin{pmatrix} x_{t|t} \\ \eta_t \end{pmatrix} \sim \mathrm{CSN}(\mu_A,\ \Sigma_A,\ \Gamma_A,\ \nu_A,\ \Delta_A)$$

where

$$\mu_A = A\mu_j = \begin{bmatrix} \mu_{t|t} \\ G\mu_{t|t} + \mu_\eta \end{bmatrix}$$

$$\Sigma_A = A\Sigma_j A' = \begin{bmatrix} \Sigma_{t|t} & \Sigma_{t|t}G' \\ G\Sigma_{t|t} & G\Sigma_{t|t}G' + \Sigma_\eta \end{bmatrix}$$

$$\Gamma_A = \Gamma_j A^{-1} = \begin{bmatrix} \Gamma_{t|t} & 0 \\ 0 & \Gamma_\eta \end{bmatrix} \begin{bmatrix} I & 0 \\ -G & I \end{bmatrix} = \begin{bmatrix} \Gamma_{t|t} & 0 \\ -\Gamma_\eta G & \Gamma_\eta \end{bmatrix}$$

$$\nu_A = \nu_j = \begin{bmatrix} \nu_{t|t} \\ \nu_\eta \end{bmatrix}$$

$$\Delta_A = \Delta_j = \begin{bmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{bmatrix}$$

Using the rules for conditional distributions (property 5 of the paper) we obtain

$$x_t|x_{t+1}, \mathcal{F}_t \sim CSN(\mu_{dt}, \Sigma_{dt}, \Gamma_{dt}, \nu_{dt}, \Delta_{dt}).$$

□

Using the notations above and some simplifications, we can rewrite out theorem as follows:

$$x_t|x_{t+1}, \mathcal{F}_t \sim CSN(\mu_{dt}, \Sigma_{dt}, \Gamma_{dt}, \nu_{dt}, \Delta_{dt})$$

with

$$\mu_{dt} = \mu_{t|t} + J_t(x_{t+1} - \mu_{t+1|t})$$

$$\Sigma_{dt} = \Sigma_{t|t} - J_t G \Sigma_{t|t}$$

$$\Gamma_{dt} = \begin{pmatrix} \Gamma_{t|t} \\ -\Gamma_\eta G \end{pmatrix}$$

$$\nu_{dt} = \nu_{t+1|t} - \Gamma_{t+1|t}(x_{t+1} - \mu_{t+1|t})$$

$$\Delta_{dt} = \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{pmatrix}.$$

Theorem 2. *The following equality holds,*

$$\Phi(\Gamma_{t+1|t+1}(x_{t+1} - \mu_{t+1|t+1}); \nu_{t+1|t+1}, \Delta_{t+1|t+1}) = \Phi(0; \nu_{dt}, \Delta_{dt} + \Gamma_{dt}\Sigma_{dt}\Gamma'_{dt})$$

or, equivalently,

$$\Phi(0; \nu_{t+1|t+1} - \Gamma_{t+1|t+1}(x_{t+1} - \mu_{t+1|t+1}), \Delta_{t+1|t+1}) = \Phi(0; \nu_{dt}, \Delta_{dt} + \Gamma_{dt}\Sigma_{dt}\Gamma'_{dt}).$$

Proof. We rewrite each expression in more basic terms. For the left hand side we obtain

$$\begin{aligned}
\Gamma_{t+1|t+1}(x_{t+1} - \mu_{t+1|t+1}) &= \Gamma_{t+1|t}(x_{t+1} - \mu_{t+1|t} - \Sigma_{t+1|t}F'(F\Sigma_{t+1|t}F' + \Sigma_\varepsilon)^{-1}(y_{t+1} - F\mu_{t+1|t})) \\
\nu_{t+1|t+1} &= \nu_{t+1|t} - \Gamma_{t+1|t}\Sigma_{t+1|t}F'(F\Sigma_{t+1|t}F' + \Sigma_\varepsilon)^{-1}(y_{t+1} - F\mu_{t+1|t})
\end{aligned}$$

and

$$\Delta_{t+1|t+1} = \begin{pmatrix} \Delta_{t|t} + \Gamma_{t|t}\Sigma_{t|t}\Gamma'_{t|t} - \Gamma_{t|t}\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}G\Sigma_{t|t}\Gamma'_{t|t} & -\Gamma_{t|t}\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}\Sigma_\eta\Gamma'_\eta \\ (-\Gamma_{t|t}\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}\Sigma_\eta\Gamma'_\eta)' & \Delta_\eta + \Gamma_\eta\Sigma_\eta\Gamma'_\eta - \Gamma_\eta\Sigma_\eta\Sigma^{-1}_{t+1|t}\Sigma_\eta\Gamma'_\eta \end{pmatrix}.$$

Hence,

$$\begin{aligned}
\nu_{t+1|t+1} - \Gamma_{t+1|t+1}(x_{t+1} - \mu_{t+1|t+1}) &= \nu_{t+1|t} - \Gamma_{t+1|t}(x_{t+1} - \mu_{t+1|t}) \\
&= \begin{pmatrix} \nu_{t|t} \\ \nu_\eta \end{pmatrix} - \begin{pmatrix} \Gamma_{t|t}\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t} \\ \Gamma_\eta\Sigma_\eta\Sigma^{-1}_{t+1|t} \end{pmatrix}(x_{t+1} - \mu_{t+1|t}) \\
&= \begin{pmatrix} \nu_{t|t} - \Gamma_{t|t}\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}(x_{t+1} - \mu_{t+1|t}) \\ \nu_\eta - \Gamma_\eta\Sigma_\eta\Sigma^{-1}_{t+1|t}(x_{t+1} - \mu_{t+1|t}) \end{pmatrix}.
\end{aligned}$$

For the right hand side we obtain

$$\begin{aligned}
\nu_{dt} &= \begin{pmatrix} \nu_{t|t} - \Gamma_{t|t}\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}(x_{t+1} - G\mu_{t|t} - \mu_\eta) \\ \nu_\eta - \Gamma_\eta + \Gamma_\eta G\Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}(x_{t+1} - G\mu_{t|t} - \mu_\eta) \end{pmatrix} \\
\Delta_{dt} + \Gamma_{dt}\Sigma_{dt}\Gamma'_{dt} &= \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{pmatrix} + \begin{pmatrix} \Gamma_{t|t} \\ -\Gamma_\eta G \end{pmatrix}(\Sigma_{t|t} - \Sigma_{t|t}G'\Sigma^{-1}_{t+1|t}G\Sigma_{t|t})\begin{pmatrix} \Gamma_{t|t} \\ -\Gamma_\eta G \end{pmatrix}'
\end{aligned}$$

Closer inspection shows that both expressions are the same.

$\square$

Theorem 3.

$$\nu^{top}_{T|T} = \nu_{t+1|t} - \Gamma_{t+1|t}(\mu_{t+1|T} - \mu_{t+1|t}) \tag{1}$$

where $\nu^{top}_{T|T}$ are the top $t+2$ elements of $\nu_{T|T}$ such that the dimensions fit.

Proof. Note that $K_t = \mu_{t+1|t+1} - \mu_{t+1|t}$.

For period $T-1$: It can be easily seen that:

$$\nu_{T|T} = \nu_{T|T-1} - \Gamma_{T|T-1}(\mu_{T|T} - \mu_{T|T-1}).$$

For period $T-2$: We start with the equation above

$$\begin{aligned}
\nu^{top}_{T|T} &= \nu_{T|T-1} - \Gamma_{T|T-1}(\mu_{T|T} - \mu_{T|T-1}) \\
&= \begin{pmatrix} \nu_{T-1|T-1} \\ \nu_\eta \end{pmatrix} - \begin{pmatrix} \Gamma_{T-1|T-1}J_{T-1} \\ \Gamma_\eta \Sigma_\eta \Sigma^{-1}_{T|T-1} \end{pmatrix} (\mu_{T|T} - \mu_{T|T-1}).
\end{aligned}$$

Take the first row and the corresponding $\nu^{top}_{T|T}$:

$$\begin{aligned}
\nu^{top}_{T|T} &= \nu_{T-1|T-1} - \Gamma_{T-1|T-1}J_{T-1}(\mu_{T|T} - \mu_{T|T-1}) \\
&= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}K_{T-2} - \Gamma_{T-1|T-2}J_{T-1}(\mu_{T|T} - \mu_{T|T-1}) \\
&= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}\big[K_{T-2} + J_{T-1}(\mu_{T|T} - \mu_{T|T-1})\big] \\
&= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}\big[\mu_{T-1|T-1} - \mu_{T-1|T-2} + J_{T-1}(\mu_{T|T} - \mu_{T|T-1})\big] \\
&= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}\big[\mu_{T-1|T} - \mu_{T-1|T-2}\big].
\end{aligned}$$

For period $T-3$: We start with

$$\begin{aligned}
\nu^{top}_{T|T} &= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}(\mu_{T-1|T} - \mu_{T-1|T-2}) \\
&= \begin{pmatrix} \nu_{T-2|T-2} \\ \nu_\eta \end{pmatrix} - \begin{pmatrix} \Gamma_{T-2|T-2}J_{T-2} \\ \Gamma_\eta \Sigma_\eta \Sigma^{-1}_{T-1|T-2} \end{pmatrix} (\mu_{T-1|T} - \mu_{T-1|T-2}).
\end{aligned}$$

Again take the first row and the corresponding $\nu^{top}_{T|T}$:

$$
\begin{aligned}
\nu^{top}_{T|T} &= \nu_{T-2|T-2} - \Gamma_{T-2|T-2} J_{T-1}(\mu_{T-1|T} - \mu_{T-1|T-2}) \\
&= \nu_{T-2|T-3} - \Gamma_{T-2|T-3} K_{T-3} - \Gamma_{T-2|T-3} J_{T-2}(\mu_{T-1|T} - \mu_{T-1|T-2}) \\
&= \nu_{T-2|T-3} - \Gamma_{T-2|T-3}\big[K_{T-3} + J_{T-2}(\mu_{T-1|T} - \mu_{T-1|T-2})\big] \\
&= \nu_{T-2|T-3} - \Gamma_{T-2|T-3}\big[\mu_{T-2|T-2} - \mu_{T-2|T-3} + J_{T-2}(\mu_{T-1|T} - \mu_{T-1|T-2})\big] \\
&= \nu_{T-2|T-3} - \Gamma_{T-2|T-3}\big[\mu_{T-2|T} - \mu_{T-2|T-3}\big].
\end{aligned}
$$

For any period t: Assume the following holds for $t+1$:

$$
\nu^{top}_{T|T} = \nu_{t+2|t+1} - \Gamma_{t+2|t+1}(\mu_{t+2|T} - \mu_{t+2|t+1}).
$$

Then

$$
\begin{aligned}
\nu^{top}_{T|T} &= \nu_{t+2|t+1} - \Gamma_{t+2|t+1}(\mu_{t+2|T} - \mu_{t+2|t+1}) \\
&= \begin{pmatrix} \nu_{t+1|t+1} \\ \nu_\eta \end{pmatrix} - \begin{pmatrix} \Gamma_{t+1|t+1} J_{t+1} \\ \Gamma_\eta \Sigma_\eta \Sigma^{-1}_{t+2|t+1} \end{pmatrix} (\mu_{t+2|T} - \mu_{t+2|t+1}).
\end{aligned}
$$

Again take the first row and the corresponding $\nu^{top}_{T|T}$:

$$
\begin{aligned}
\nu^{top}_{T|T} &= \nu_{t+1|t+1} - \Gamma_{t+1|t+1} J_{t+1}(\mu_{t+2|T} - \mu_{t+2|t+1}) \\
&= \nu_{t+1|t} - \Gamma_{t+1|t} K_t - \Gamma_{t+1|t} J_{t+1}(\mu_{t+2|T} - \mu_{t+2|t+1}) \\
&= \nu_{t+1|t} - \Gamma_{t+1|t}\big[K_t + J_{t+1}(\mu_{t+2|T} - \mu_{t+2|t+1})\big] \\
&= \nu_{t+1|t} - \Gamma_{t+1|t}\big[\mu_{t+1|t+1} - \mu_{t+1|t} + J_{t+1}(\mu_{t+2|T} - \mu_{t+2|t+1})\big] \\
&= \nu_{t+1|t} - \Gamma_{t+1|t}\big[\mu_{t+1|T} - \mu_{t+1|t}\big].
\end{aligned}
$$

$\square$

Lemma 1. *For any t, the following holds:*

$$\Phi\Big[\Gamma_{dt}(x_t-\mu_{dt});\nu_{dt},\Delta_{dt}\Big]=\Phi\Bigg[\begin{pmatrix}\Gamma_{t|t} & 0\\ -\Gamma_\eta G & \Gamma_\eta\end{pmatrix}\left(\begin{pmatrix}x_t\\ x_{t+1}\end{pmatrix}-\begin{pmatrix}\mu_{t|T}\\ \mu_{t+1|T}\end{pmatrix}\right); \\ \nu_{t+1|t}-\Gamma_{t+1|t}(\mu_{t+1|T}-\mu_{t+1|t}), \\ \begin{pmatrix}\Delta_{t|t} & 0\\ 0 & \Delta_\eta\end{pmatrix}\Bigg] \tag{2}$$

where

$$\mu_{t|T}\equiv\mu_{t|t}+\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}(\mu_{t+1|T}-\mu_{t+1|t})$$

which coincides with the first parameter of the distribution of $x_{t|T}$.

Proof. Note that $\Gamma_\eta GJ_t-\Gamma_\eta=-\Gamma_\eta\Sigma_\eta\Sigma_{t+1|t}^{-1}$.

$$\begin{aligned}
&\Phi\left[\Gamma_{dt}(x_t-\mu_{dt});\nu_{dt},\Delta_{dt}\right]\\
&=\Phi\left[\begin{pmatrix}\Gamma_{t|t}\\ -\Gamma_\eta G\end{pmatrix}(x_t-\mu_{t|t})-\begin{pmatrix}\Gamma_{t|t}J_t\\ -\Gamma_\eta GJ_t\end{pmatrix}(x_{t+1}-\mu_{t+1|t});\right.\\
&\qquad\left.\nu_{t+1|t}-\begin{pmatrix}\Gamma_{t|t}J_t\\ \Gamma_\eta\Sigma_\eta\Sigma_{t+1|t}^{-1}\end{pmatrix}(x_{t+1}-\mu_{t+1|t}),\begin{pmatrix}\Delta_{t|t} & 0\\ 0 & \Delta_\eta\end{pmatrix}\right]\\
&=\Phi\left[\Gamma_{t|t}(x_t-\mu_{t|t});\nu_{t|t},\Delta_{t|t}\right]\times\Phi\left[-\Gamma_\eta G(x_t-\mu_{t|t})+\Gamma_\eta(x_{t+1}-\mu_{t+1|t});\nu_\eta,\Delta_\eta\right]
\end{aligned}$$

Expand these cdfs

$$
\begin{aligned}
&\Phi\left[\Gamma_{t|t}(x_t-\mu_{t|t});\nu_{t|t},\Delta_{t|t}\right]\times\Phi\left[-\Gamma_\eta G(x_t-\mu_{t|t})+\Gamma_\eta(x_{t+1}-\mu_{t+1|t});\nu_\eta,\Delta_\eta\right]\\
&=\Phi\left[\Gamma_{t|t}(x_t-\mu_{t|t}-J_t(\mu_{t+1|T}-\mu_{t+1|t}));\ \nu_{t|t}-\Gamma_{t|t}J_t(\mu_{t+1|T}-\mu_{t+1|t}),\ \Delta_{t|t}\right]\\
&\quad\times\Phi\big[-\Gamma_\eta G(x_t-\mu_{t|t}-J_t(\mu_{t+1|T}-\mu_{t+1|t})+J_t(\mu_{t+1|T}-\mu_{t+1|t}))\\
&\qquad\qquad+\Gamma_\eta(x_{t+1}-\mu_{t+1|T}+\mu_{t+1|T}-\mu_{t+1|t});\ \nu_\eta,\ \Delta_\eta\big]\\
&=\Phi\left[\Gamma_{t|t}(x_t-\mu_{t|T});\ \nu_{t|t}-\Gamma_{t|t}J_t(\mu_{t+1|T}-\mu_{t+1|t}),\ \Delta_{t|t}\right]\\
&\quad\times\Phi\big[-\Gamma_\eta G(x_t-\mu_{t|T})+\Gamma_\eta(x_{t+1}-\mu_{t+1|T})\\
&\qquad\qquad-\Gamma_\eta GJ_t(\mu_{t+1|T}-\mu_{t+1|t})+\Gamma_\eta(\mu_{t+1|T}-\mu_{t+1|t});\ \nu_\eta,\ \Delta_\eta\big]\\
&=\Phi\left[\Gamma_{t|t}(x_t-\mu_{t|T});\ \nu_{t|t}-\Gamma_{t|t}J_t(\mu_{t+1|T}-\mu_{t+1|t}),\ \Delta_{t|t}\right]\\
&\quad\times\Phi\left[-\Gamma_\eta G(x_t-\mu_{t|T})+\Gamma_\eta(x_{t+1}-\mu_{t+1|T});\ \nu_\eta+(\Gamma_\eta GJ_t-\Gamma_\eta)(\mu_{t+1|T}-\mu_{t+1|t}),\ \Delta_\eta\right]\\
&=\Phi\left[\begin{pmatrix}\Gamma_{t|t} & 0\\ -\Gamma_\eta G & \Gamma_\eta\end{pmatrix}\left(\begin{pmatrix}x_t\\ x_{t+1}\end{pmatrix}-\begin{pmatrix}\mu_{t|T}\\ \mu_{t+1|T}\end{pmatrix}\right);\nu_{t+1|t}-\Gamma_{t+1|t}(\mu_{t+1|T}-\mu_{t+1|t}),\begin{pmatrix}\Delta_{t|t} & 0\\ 0 & \Delta_\eta\end{pmatrix}\right]
\end{aligned}
$$

□

Lemma 2. *As to the product of two normal pdfs, we obtain for any t*

$$
\begin{aligned}
\phi\left[x_t;\ \mu_{dt},\ \Sigma_{dt}\right]\times\phi\left[x_{t+1};\ \mu_{t+1|T},\ \Sigma_{t+1|T}\right]=\phi\Bigg[&\begin{pmatrix}x_t\\ x_{t+1}\end{pmatrix};\\
&\begin{pmatrix}\mu_{t|t}+J_t(\mu_{t+1|T}-\mu_{t+1|t})\\ \mu_{t+1|T}\end{pmatrix},\\
&\begin{pmatrix}\Sigma_{t|t}+J_t(\Sigma_{t+1|T}-\Sigma_{t+1|t})J_t' & J_t\Sigma_{t+1|T}\\ \Sigma_{t+1|T}J_t' & \Sigma_{t+1|T}\end{pmatrix}\Bigg]
\end{aligned}
\tag{3}
$$

where

$$
\begin{aligned}
\mu_{t|T}&\equiv\mu_{t|t}+\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}(\mu_{t+1|T}-\mu_{t+1|t})\\
\Sigma_{t|T}&\equiv\Sigma_{t|t}+\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}(\Sigma_{t+1|T}-\Sigma_{t+1|t})\Sigma_{t+1|t}^{-1}G\Sigma_{t|t}
\end{aligned}
$$

which coincides with the first two parameters of the distribution of $x_{t|T}$.

Proof. Recall the conditioning property of the normal distribution. Let $X\sim N_p(\psi,\Omega)$ be partitioned into

X_1 of length p_1 and X_2 of length p_2, such that $X = (X_1', X_2')'$. The parameters are partitioned accordingly,

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}.$$

Then

$$X_{1|2} = (X_1 | X_2 = x_2) \sim N_{p_1}(\psi_{1|2}, \Omega_{1|2})$$

with $\psi_{1|2} = \psi_1 + \Omega_{12}\Omega_{22}^{-1}(x_2 - \psi_2)$ and, $\Omega_{1|2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$. If the marginal distribution of X_2 is singular, the inverse Ω_{22}^{-1} has to be replaced by the pseudo-inverse.

Write the left hand side of 3 in more basic terms and add and subtract some of its expressions

$$\begin{aligned}
&\phi\left[x_t;\ \mu_{t|t} + J_t(x_{t+1} - \mu_{t+1|t}),\ \Sigma_{t|t} - J_t G \Sigma_{t|t}\right] \times \phi\left[x_{t+1};\ \mu_{t+1|T},\ \Sigma_{t+1|T}\right] \\
&= \phi\Big[x_t; \\
&\qquad \mu_{t|t} + J_t \Sigma_{t+1|t} \Sigma_{t+1|t}^{-1}(x_{t+1} - \mu_{t+1|T} + \mu_{t+1|T} - \mu_{t+1|t}), \\
&\qquad \Sigma_{t|t} - J_t \Sigma_{t+1|t} \underbrace{\Sigma_{t+1|t}^{-1} G \Sigma_{t|t}}_{=J_t'} + J_t \Sigma_{t+1|T} \Sigma_{t+1|T}^{-1} \Sigma_{t+1|T} J_t' - \underbrace{J_t \Sigma_{t+1|T}}_{\text{analogous to } \Omega_{12}} \Sigma_{t+1|T}^{-1} \Sigma_{t+1|T} J_t'\Big] \\
&\quad \times \phi\left[x_{t+1};\ \mu_{t+1|T},\ \Sigma_{t+1|T}\right] \\
&= \phi\Big[x_t; \\
&\qquad \mu_{t|t} + J_t(\mu_{t+1|T} - \mu_{t+1|t}) + \underbrace{J_t \Sigma_{t+1|t}}_{\text{analogous to } \Omega_{12}} \underbrace{\Sigma_{t+1|t}^{-1}}_{\Omega_{22}^{-1}}(x_{t+1} - \mu_{t+1|T}), \\
&\qquad \Sigma_{t|t} - J_t \Sigma_{t+1|t} J_t' + J_t \Sigma_{t+1|T} J_t' - J_t \Sigma_{t+1|T} \Sigma_{t+1|T}^{-1} \Sigma_{t+1|T} J_t'\Big] \\
&\quad \times \phi\left[x_{t+1};\ \mu_{t+1|T},\ \Sigma_{t+1|T}\right].
\end{aligned}$$

Using the conditioning property (in reverse), equation (3) follows. □

Lemma 3.

$$\Delta_{dt} + \Gamma_{dt} \Sigma_{dt} \Gamma_{dt}' = \Delta_{t+1|t+1}. \tag{4}$$

Proof. Write $\Delta_{dt} + \Gamma_{dt}\Sigma_{dt}\Gamma'_{dt}$ in more basic terms and evaluate:

$$
\begin{aligned}
\Delta_{dt} + \Gamma_{dt}\Sigma_{dt}\Gamma'_{dt} &= \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{pmatrix} + \begin{pmatrix} \Gamma_{t|t} \\ -\Gamma_\eta G \end{pmatrix} (\Sigma_{t|t} - J_t G \Sigma_{t|t}) \begin{pmatrix} \Gamma'_{t|t} & -G'\Gamma'_\eta \end{pmatrix} \\
&= \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \Delta_\eta \end{pmatrix} + \begin{pmatrix} \Gamma_{t|t}(\Sigma_{t|t} - J_t G\Sigma_{t|t})\Gamma'_{t|t} & -\Gamma_{t|t}(\Sigma_{t|t} - J_t G\Sigma_{t|t})G'\Gamma'_\eta \\ -\Gamma_\eta G(\Sigma_{t|t} - J_t G\Sigma_{t|t})\Gamma'_{t|t} & \Gamma_\eta G(\Sigma_{t|t} - J_t G\Sigma_{t|t})G'\Gamma'_\eta \end{pmatrix} \\
&= \Delta_{t+1|t} \\
&= \Delta_{t+1|t+1},
\end{aligned}
$$

where we used $G\Sigma_{t|t}G' = \Sigma_{t+1|t} - \Sigma_\eta$. □

3.3. Smoothing formulas for period T-1

In the final period, $x_{T|T}$ is both the filtered and the smoothed distribution. This is where the backward recursion kicks in. Consider the penultimate period $T-1$. The conditional density of $x_{T-1}|x_T, \mathcal{F}_{T-1}$ is (see theorem 1)

$$
\frac{\Phi(\Gamma_{dT-1}(x_{T-1} - \mu_{dT-1}); \nu_{dT-1}, \Delta_{dT-1})}{\Phi(0; \nu_{dT-1}, \Delta_{dT-1} + \Gamma_{dT-1}\Sigma_{dT-1}\Gamma'_{dT-1})} \phi(x_{T-1}; \mu_{dT-1}, \Sigma_{dT-1})
$$

where both μ_{dT-1} and ν_{dT-1} are functions of x_T. To find the distribution of $x_{T-1}|\mathcal{F}_T$ we average the density of $x_{T-1}|x_T, \mathcal{F}_{T-1}$ over $x_T|\mathcal{F}_T$,

$$
\begin{aligned}
\int \frac{\Phi(\Gamma_{dT-1}(x_{T-1} - \mu_{dT-1}); \nu_{dT-1}, \Delta_{dT-1})}{\Phi(0; \nu_{dT-1}, \Delta_{dT-1} + \Gamma_{dT-1}\Sigma_{dT-1}\Gamma'_{dT-1})} \frac{\Phi(\Gamma_{T|T}(x_T - \mu_{T|T}); \nu_{T|T}, \Delta_{T|T})}{\Phi(0; \nu_{T|T}, \Delta_{T|T} + \Gamma_{T|T}\Sigma_{T|T}\Gamma'_{T|T})} \\
\times \phi(x_{T-1}; \mu_{dT-1}, \Sigma_{dT-1})\phi(x_T; \mu_{T|T}, \Sigma_{T|T}) dx_T.
\end{aligned}
$$

Due to theorem 2, the top right cdf and the bottom left cdf are identical,

$$
\Phi(\Gamma_{T|T}(x_T - \mu_{T|T}); \nu_{T|T}, \Delta_{T|T}) = \Phi(0; \nu_{dT-1}, \Delta_{dT-1} + \Gamma_{dT-1}\Sigma_{dT-1}\Gamma'_{dT-1}).
$$

Having canceled the cdfs, the integral simplifies to

$$
\int \frac{\Phi(\Gamma_{dT-1}(x_{T-1} - \mu_{dT-1}); \nu_{dT-1}, \Delta_{dT-1})}{\Phi(0; \nu_{T|T}, \Delta_{T|T} + \Gamma_{T|T}\Sigma_{T|T}\Gamma'_{T|T})} \times \phi(x_{T-1}; \mu_{dT-1}, \Sigma_{dT-1})\phi(x_T; \mu_{T|T}, \Sigma_{T|T}) dx_T.
$$

We first look at the joint distribution of x_{T-1} and x_T (given $\mathcal{F}_T$) and then marginalize. Let μ_{jT-1}, Σ_{jT-1},

Γ_{jT-1}, ν_{jT-1}, Δ_{jT-1} denote the parameters of the joint distribution. Obviously, by lemma 2

$$
\begin{aligned}
\mu_{jT-1} &= \begin{pmatrix} \mu_{T-1|T-1} + J_{T-1}(\mu_{T|T} - \mu_{T|T-1}) \\ \mu_{T|T} \end{pmatrix} \\
\Sigma_{jT-1} &= \begin{pmatrix} \Sigma_{T-1|T-1} + J_{T-1}(\Sigma_{T|T} - \Sigma_{T|T-1})J'_{T-1} & J_{T-1}\Sigma_{T|T} \\ \Sigma_{T|T}J'_{T-1} & \Sigma_{T|T} \end{pmatrix}.
\end{aligned}
$$

Less obvious, but still relatively straightforward (use lemma 1 and lemma 3),

$$
\begin{aligned}
\Gamma_{jT-1} &= \begin{pmatrix} \Gamma_{T-1|T-1} & 0 \\ -\Gamma_\eta G & \Gamma_\eta \end{pmatrix} \\
\nu_{jT-1} &= \nu_{T|T} \\
\Delta_{jT-1} &= \begin{pmatrix} \Delta_{T-1|T-1} & 0 \\ 0 & \Delta_\eta \end{pmatrix}.
\end{aligned}
$$

Using lemma 2.3.1 of Genton (2004), the marginal distribution of $x_{T-1}|\mathcal{F}_T$ is

$$
\begin{aligned}
\mu_{T-1|T} &= \mu_{T-1|T-1} + J_{T-1}(\mu_{T|T} - \mu_{T|T-1}) \\
\Sigma_{T-1|T} &= \Sigma_{T-1|T-1} + J_{T-1}(\Sigma_{T|T} - \Sigma_{T|T-1})J'_{T-1} \\
\Gamma_{T-1|T} &= \begin{pmatrix} \Gamma_{T-1|T-1} \\ N_{T-1} \end{pmatrix} \\
\nu_{T-1|T} &= \nu_{T|T} \\
\Delta_{T-1|T} &= \begin{pmatrix} \Delta_{T-1|T-1} & 0 \\ 0 & \tilde{\Delta}_{T-1} \end{pmatrix}
\end{aligned}
$$

where

$$
\begin{aligned}
N_{T-1} &= -\Gamma_\eta G + \Gamma_\eta M_{T-1} \\
M_{T-1} &= \Sigma_{T|T}J'_{T-1}\Sigma^{-1}_{T-1|T} \\
L_{T-1} &= \Sigma_{T|T} - M_{T-1}\Sigma_{T-1|T}M'_{T-1}
\end{aligned}
$$

and

$$
\tilde{\Delta}_{T-1} = \Delta_\eta + \Gamma_\eta L_{T-1}\Gamma'_\eta.
$$

3.4. Smoothing formulas for period T-2

The conditional density of $x_{T-2}|x_{T-1}, \mathcal{F}_{T-2}$ is

$$\frac{\Phi(\Gamma_{dT-2}(x_{T-2}-\mu_{dT-2});\nu_{dT-2},\Delta_{dT-2})}{\Phi(0;\nu_{dT-2},\Delta_{dT-2}+\Gamma_{dT-2}\Sigma_{dT-2}\Gamma'_{dT-2})}\phi(x_{T-2};\mu_{dT-2},\Sigma_{dT-2})$$

where μ_{dT-2} and ν_{dT-2} are functions of x_{T-1}. To find the distribution of $x_{T-2}|\mathcal{F}_T$ we again derive the joint distribution of x_{T-2} and x_{T-1} given $\mathcal{F}_T$,

$$\begin{aligned}&\frac{\Phi(\Gamma_{dT-2}(x_{T-2}-\mu_{dT-2});\nu_{dT-2},\Delta_{dT-2})}{\Phi(0;\nu_{dT-2},\Delta_{dT-2}+\Gamma_{dT-2}\Sigma_{dT-2}\Gamma'_{dT-2})}\frac{\Phi(\Gamma_{T-1|T}(x_{T-1}-\mu_{T-1|T});\nu_{T-1|T},\Delta_{T-1|T})}{\Phi(0;\nu_{T-1|T},\Delta_{T-1|T}+\Gamma_{T-1|T}\Sigma_{T-1|T}\Gamma'_{T-1|T})}\\ &\qquad\times\phi(x_{T-2};\mu_{dT-2},\Sigma_{dT-2})\phi(x_{T-1};\mu_{T-1|T},\Sigma_{T-1|T}).\end{aligned}$$

According to theorem 3 and lemma 1, for $t = T-2$, the top left cdf can be rewritten as

$$\begin{aligned}&\Phi(\Gamma_{dT-2}(x_{T-2}-\mu_{dT-2});\nu_{dT-2},\Delta_{dT-2})\\ &=\Phi\left(\begin{pmatrix}\Gamma_{T-2|T-2} & 0\\ -\Gamma_\eta G & \Gamma_\eta\end{pmatrix}\left(\begin{pmatrix}x_{T-2}\\ x_{T-1}\end{pmatrix}-\begin{pmatrix}\mu_{T-2|T}\\ \mu_{T-1|T}\end{pmatrix}\right);\nu^{top}_{T|T},\begin{pmatrix}\Delta_{T-2|T-2} & 0\\ 0 & \Delta_\eta\end{pmatrix}\right)\end{aligned}$$

The top right cdf can be factorized as

$$\begin{aligned}&\Phi(\Gamma_{T-1|T}(x_{T-1}-\mu_{T-1|T});\nu_{T-1|T},\Delta_{T-1|T})\\ &=\Phi\left(\begin{pmatrix}\Gamma_{T-1|T-1}\\ N_{T-1}\end{pmatrix}(x_{T-1}-\mu_{T-1|T});\nu_{T|T},\begin{pmatrix}\Delta_{T-1|T-1} & 0\\ 0 & \tilde{\Delta}_{T-1}\end{pmatrix}\right)\\ &=\Phi(\Gamma_{T-1|T-1}(x_{T-1}-\mu_{T-1|T});\nu^{top}_{T|T},\Delta_{T-1|T-1})\\ &\quad\times\Phi(N_{T-1}(x_{T-1}-\mu_{T-1|T});\nu^{btm}_{T|T},\tilde{\Delta}_{T-1})\end{aligned} \tag{5}$$

where $\nu_{T|T}$ is suitably partitioned into $\nu^{top}_{T|T}$ and $\nu^{btm}_{T|T}$.

Using theorem 2 and theorem 3 the bottom left cdf is

$$\begin{aligned}&\Phi(0;\nu_{dT-2},\Delta_{dT-2}+\Gamma_{dT-2}\Sigma_{dT-2}\Gamma'_{dT-2})\\ &=\Phi(\Gamma_{T-1|T-1}(x_{T-1}-\mu_{T-1|T-1});\nu_{T-1|T-1},\Delta_{T-1|T-1})\\ &=\Phi(\Gamma_{T-1|T-1}(x_{T-1}-\mu_{T-1|T}+\mu_{T-1|T}-\mu_{T-1|T-1});\nu_{T-1|T-1},\Delta_{T-1|T-1})\\ &=\Phi(\Gamma_{T-1|T-1}(x_{T-1}-\mu_{T-1|T});\nu_{T-1|T-1}-\Gamma_{T-1|T-1}(\mu_{T-1|T}-\mu_{T-1|T-1}),\Delta_{T-1|T-1})\end{aligned}$$

and

$$
\begin{aligned}
&\nu_{T-1|T-1} - \Gamma_{T-1|T-1}(\mu_{T-1|T} - \mu_{T-1|T-1}) \\
&= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}K_{T-2} - \Gamma_{T-1|T-2}(\mu_{T-1|T} - \mu_{T-1|T-2} - K_{T-2}) \\
&= \nu_{T-1|T-2} - \Gamma_{T-1|T-2}(\mu_{T-1|T} - \mu_{T-1|T-2}) \\
&= \nu_{T|T}^{top}.
\end{aligned}
$$

As a result, we have

$$
\begin{aligned}
&\Phi(0; \nu_{dT-2}, \Delta_{dT-2} + \Gamma_{dT-2}\Sigma_{dT-2}\Gamma'_{dT-2}) \\
&= \Phi(\Gamma_{T-1|T-1}(x_{T-1} - \mu_{T-1|T}); \nu_{T|T}^{top}, \Delta_{T-1|T-1})
\end{aligned}
$$

Therefore, the first part of (5) and the bottom left cdf cancel each other out. The remaining part of the top right cdf can be merged into the top left cdf. This results in the numerator

$$
\Phi\left(\begin{pmatrix} \Gamma_{T-2|T-2} & 0 \\ -\Gamma_\eta G & \Gamma_\eta \\ 0 & N_{T-1} \end{pmatrix} \begin{pmatrix} x_{T-2} - \mu_{T-2|T} \\ x_{T-1} - \mu_{T-1|T} \end{pmatrix}; \nu_{T|T}, \begin{pmatrix} \Delta_{T-2|T-2} & 0 & 0 \\ 0 & \Delta_\eta & 0 \\ 0 & 0 & \tilde{\Delta}_{T-1} \end{pmatrix}\right)
$$

Let μ_{jT-2} etc. denote the parameters of the joint distribution,

$$
\begin{aligned}
\mu_{jT-2} &= \begin{pmatrix} \mu_{T-2|T-2} + J_{T-2}(\mu_{T-1|T} - \mu_{T-1|T-2}) \\ \mu_{T-1|T} \end{pmatrix} \\
\Sigma_{jT-2} &= \begin{pmatrix} \Sigma_{T-2|T-2} + J_{T-2}(\Sigma_{T-1|T} - \Sigma_{T-1|T-2})J'_{T-2} & J_{T-2}\Sigma_{T-1|T} \\ \Sigma_{T-1|T}J'_{T-2} & \Sigma_{T-1|T} \end{pmatrix} \\
\Gamma_{jT-2} &= \begin{pmatrix} \Gamma_{T-2|T-2} & 0 \\ -\Gamma_\eta G & \Gamma_\eta \\ 0 & N_{T-1} \end{pmatrix} \\
\nu_{jT-2} &= \nu_{T|T} \\
\Delta_{jT-2} &= \begin{pmatrix} \Delta_{T-2|T-2} & 0 & 0 \\ 0 & \Delta_\eta & 0 \\ 0 & 0 & \tilde{\Delta}_{T-1} \end{pmatrix}.
\end{aligned}
$$

The marginal distribution of $x_{T-2}|\mathcal{F}_T$ is

$$
\begin{aligned}
\mu_{T-2|T} &= \mu_{T-2|T-2} + J_{T-2}(\mu_{T-1|T} - \mu_{T-1|T-2}) \\
\Sigma_{T-2|T} &= \Sigma_{T-2|T-2} + J_{T-2}(\Sigma_{T-1|T} - \Sigma_{T-1|T-2})J'_{T-2} \\
\Gamma_{T-2|T} &= \begin{pmatrix} \Gamma_{T-2|T-2} \\ N_{T-2} \\ N_{T-1}M_{T-2} \end{pmatrix} \\
\nu_{T-2|T} &= \nu_{T|T} \\
\Delta_{T-2|T} &= \begin{pmatrix} \Delta_{T-2|T-2} & 0 \\ 0 & \tilde{\Delta}_{T-2} \end{pmatrix}
\end{aligned}
$$

with

$$
\begin{aligned}
N_{T-2} &= -\Gamma_\eta G + \Gamma_\eta M_{T-2} \\
M_{T-2} &= \Sigma_{T-1|T} J'_{T-2} \Sigma^{-1}_{T-2|T} \\
L_{T-2} &= \Sigma_{T-1|T} - M_{T-2}\Sigma_{T-2|T}M'_{T-2}
\end{aligned}
$$

and

$$
\tilde{\Delta}_{T-2} = \begin{pmatrix} \Delta_\eta & 0 \\ 0 & \tilde{\Delta}_{T-1} \end{pmatrix} + \begin{pmatrix} \Gamma_\eta \\ N_{T-1} \end{pmatrix} L_{T-2} \begin{pmatrix} \Gamma_\eta \\ N_{T-1} \end{pmatrix}'.
$$

3.5. Smoothing formulas for period T-3

The conditional density of $x_{T-3}|x_{T-2}, \mathcal{F}_{T-3}$ is

$$
\frac{\Phi(\Gamma_{dT-3}(x_{T-3} - \mu_{dT-3}); \nu_{dT-3}, \Delta_{dT-3})}{\Phi(0; \nu_{dT-3}, \Delta_{dT-3} + \Gamma_{dT-3}\Sigma_{dT-3}\Gamma'_{dT-3})} \phi(x_{T-3}; \mu_{dT-3}, \Sigma_{dT-3})
$$

where μ_{dT-3} and ν_{dT-3} are functions of x_{T-2}. To find the distribution of $x_{T-3}|\mathcal{F}_T$ we again derive the joint distribution of x_{T-3} and x_{T-2} given $\mathcal{F}_T$,

$$
\begin{aligned}
&\frac{\Phi(\Gamma_{dT-3}(x_{T-3} - \mu_{dT-3}); \nu_{dT-3}, \Delta_{dT-3})}{\Phi(0; \nu_{dT-3}, \Delta_{dT-3} + \Gamma_{dT-3}\Sigma_{dT-3}\Gamma'_{dT-3})} \frac{\Phi(\Gamma_{T-2|T}(x_{T-2} - \mu_{T-2|T}); \nu_{T-2|T}, \Delta_{T-2|T})}{\Phi(0; \nu_{T-2|T}, \Delta_{T-2|T} + \Gamma_{T-2|T}\Sigma_{T-2|T}\Gamma'_{T-2|T})} \\
&\qquad \times \phi(x_{T-3}; \mu_{dT-3}, \Sigma_{dT-3})\phi(x_{T-2}; \mu_{T-2|T}, \Sigma_{T-2|T}).
\end{aligned}
$$

According to theorem 3 and lemma 1

$$
\begin{aligned}
&\Phi(\Gamma_{dT-3}(x_{T-3}-\mu_{dT-3});\nu_{dT-3},\Delta_{dT-3})\\
&=\Phi\left(\begin{pmatrix}\Gamma_{T-3|T-3} & 0\\ -\Gamma_\eta G & \Gamma_\eta\end{pmatrix}\left(\begin{pmatrix}x_{T-3}\\ x_{T-2}\end{pmatrix}-\begin{pmatrix}\mu_{T-3|T}\\ \mu_{T-2|T}\end{pmatrix}\right);\nu_{T|T}^{top},\begin{pmatrix}\Delta_{T-3|T-3} & 0\\ 0 & \Delta_\eta\end{pmatrix}\right).
\end{aligned}
$$

The top right cdf can be factorized as

$$
\begin{aligned}
&\Phi(\Gamma_{T-2|T}(x_{T-2}-\mu_{T-2|T});\nu_{T-2|T},\Delta_{T-2|T})\\
&=\Phi\left(\begin{pmatrix}\Gamma_{T-2|T-2}\\ N_{T-2}\\ N_{T-1}M_{T-2}\end{pmatrix}(x_{T-2}-\mu_{T-2|T});\nu_{T|T},\begin{pmatrix}\Delta_{T-2|T-2} & 0\\ 0 & \tilde{\Delta}_{T-2}\end{pmatrix}\right)\\
&=\Phi(\Gamma_{T-2|T-2}(x_{T-2}-\mu_{T-2|T});\nu_{T|T}^{top},\Delta_{T-2|T-2})\\
&\quad\times\Phi\left(\begin{pmatrix}N_{T-2}\\ N_{T-1}M_{T-2}\end{pmatrix}(x_{T-2}-\mu_{T-2|T});\nu_{T|T}^{btm},\tilde{\Delta}_{T-2}\right)
\end{aligned}
$$

where $\nu_{T|T}$ is suitably partitioned into $\nu_{T|T}^{top}$ and $\nu_{T|T}^{btm}$.

Using theorem 2 and theorem 3, the bottom left cdf is (as in the last section)

$$
\begin{aligned}
&\Phi(0;\nu_{dT-3},\Delta_{dT-3}+\Gamma_{dT-3}\Sigma_{dT-3}\Gamma'_{dT-3})\\
&=\Phi(\Gamma_{T-2|T-2}(x_{T-2}-\mu_{T-2|T});\nu_{T|T}^{top},\Delta_{T-2|T-2}).
\end{aligned}
$$

Therefore, the first part of the factorization of the top right cdf and the bottom left cdf cancel each other out. The remaining part of the top right cdf can be merged into the top left cdf. This results in the numerator

$$
\Phi\left(\begin{pmatrix}\Gamma_{T-3|T-3} & 0\\ -\Gamma_\eta G & \Gamma_\eta\\ 0 & N_{T-2}\\ 0 & N_{T-1}M_{T-2}\end{pmatrix}\begin{pmatrix}x_{T-3}-\mu_{T-3|T}\\ x_{T-2}-\mu_{T-2|T}\end{pmatrix};\nu_{T|T},\begin{pmatrix}\Delta_{T-3|T-3} & 0 & 0\\ 0 & \Delta_\eta & 0\\ 0 & 0 & \tilde{\Delta}_{T-2}\end{pmatrix}\right).
$$

Let μ_{jT-3} etc denote the parameters of the joint distribution,

$$\mu_{jT-3} = \begin{pmatrix} \mu_{T-3|T-3} + J_{T-3}(\mu_{T-2|T} - \mu_{T-2|T-3}) \\ \mu_{T-2|T} \end{pmatrix}$$

$$\Sigma_{jT-3} = \begin{pmatrix} \Sigma_{T-3|T-3} + J_{T-3}(\Sigma_{T-2|T} - \Sigma_{T-2|T-3})J'_{T-3} & J_{T-3}\Sigma_{T-2|T} \\ \Sigma_{T-2|T}J'_{T-3} & \Sigma_{T-2|T} \end{pmatrix}$$

$$\Gamma_{jT-3} = \begin{pmatrix} \Gamma_{T-3|T-3} & 0 \\ -\Gamma_\eta G & \Gamma_\eta \\ 0 & N_{T-2} \\ 0 & N_{T-1}M_{T-2} \end{pmatrix}$$

$$\nu_{jT-3} = \nu_{T|T}$$

$$\Delta_{jT-3} = \begin{pmatrix} \Delta_{T-3|T-3} & 0 & 0 \\ 0 & \Delta_\eta & 0 \\ 0 & 0 & \tilde{\Delta}_{T-2} \end{pmatrix}.$$

The marginal distribution of $x_{T-3}|\mathcal{F}_T$ is

$$\mu_{T-3|T} = \mu_{T-3|T-3} + J_{T-3}(\mu_{T-2|T} - \mu_{T-2|T-3})$$

$$\Sigma_{T-3|T} = \Sigma_{T-3|T-3} + J_{T-3}(\Sigma_{T-2|T} - \Sigma_{T-2|T-3})J'_{T-3}$$

$$\Gamma_{T-3|T} = \begin{pmatrix} \Gamma_{T-3|T-3} \\ N_{T-3} \\ N_{T-2}M_{T-3} \\ N_{T-1}M_{T-2}M_{T-3} \end{pmatrix}$$

$$\nu_{T-3|T} = \nu_{T|T}$$

$$\Delta_{T-3|T} = \begin{pmatrix} \Delta_{T-3|T-3} & 0 \\ 0 & \tilde{\Delta}_{T-3} \end{pmatrix}$$

with

$$\tilde{\Delta}_{T-3} = \begin{pmatrix} \Delta_\eta & 0 \\ 0 & \tilde{\Delta}_{T-2} \end{pmatrix} + \begin{pmatrix} \Gamma_\eta \\ N_{T-2} \\ N_{T-1}M_{T-2} \end{pmatrix} L_{T-3} \begin{pmatrix} \Gamma_\eta \\ N_{T-2} \\ N_{T-1}M_{T-2} \end{pmatrix}',$$

note that the dimension of $\tilde{\Delta}_{T-2}$ is such that the sum fits.

3.6. Smoothing formulas for any time period

The CSN parameters for $x_t|\mathcal{F}_T$ are

$$\begin{aligned}
\mu_{t|T} &= \mu_{t|t} + J_t(\mu_{t+1|T} - \mu_{t+1|t}) \\
\Sigma_{t|T} &= \Sigma_{t|t} + J_t(\Sigma_{t+1|T} - \Sigma_{t+1|t})J_t' \\
\Gamma_{t|T} &= \begin{pmatrix} \Gamma_{t|t} \\ N_t \\ N_{t+1}M_t \\ N_{t+2}M_{t+1}M_t \\ \vdots \\ N_{T-1}\cdot \ldots \cdot M_{t+2}M_{t+1}M_t \end{pmatrix} \\
\nu_{t|T} &= \nu_{T|T} \\
\Delta_{t|T} &= \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \tilde{\Delta}_t \end{pmatrix}
\end{aligned}$$

with

$$\tilde{\Delta}_t = \begin{pmatrix} \Delta_\eta & 0 \\ 0 & \tilde{\Delta}_{t+1} \end{pmatrix} + \begin{pmatrix} \Gamma_\eta \\ N_{t+1} \\ N_{t+2}M_{t+1} \\ N_{T-1}\cdot \ldots \cdot M_{t+2}M_{t+1} \end{pmatrix} L_t \begin{pmatrix} \Gamma_\eta \\ N_{t+1} \\ N_{t+2}M_{t+1} \\ N_{T-1}\cdot \ldots \cdot M_{t+2}M_{t+1} \end{pmatrix}'$$

and

$$\begin{aligned}
J_t &= \Sigma_{t|t}G'\Sigma_{t+1|t}^{-1} \\
M_t &= \Sigma_{t+1|T}J_t'\Sigma_{t|T}^{-1} \\
N_t &= -\Gamma_\eta G + \Gamma_\eta M_t \\
L_t &= \Sigma_{t+1|T} - M_t\Sigma_{t|T}M_t'.
\end{aligned}$$

We can write the formulas for general t more compactly. Abbreviate $\Sigma_{t|t}G'\Sigma_{t+1|t}^{-1}$ as J_t, $\Sigma_{t+1|T} - M_t\Sigma_{t|T}M_t'$ as L_t and define $O_{T-1} \equiv N_{T-1}$ and

$$O_t \equiv \begin{bmatrix} N_t \\ O_{t+1}M_t \end{bmatrix}$$

for $t = T-2, T-3, \ldots, 1$ (O_T is not defined). Then

$$\begin{aligned}
\mu_{t|T} &= \mu_{t|t} + \Sigma_{t|t} G' \Sigma_{t+1|t}^{-1} (\mu_{t+1|T} - \mu_{t+1|t}) \\
\Sigma_{t|T} &= \Sigma_{t|t} + \Sigma_{t|t} G' \Sigma_{t+1|t}^{-1} (\Sigma_{t+1|T} - \Sigma_{t+1|t}) \Sigma_{t+1|t}^{-1} G \Sigma_{t|t} \\
\Gamma_{t|T} &= \begin{pmatrix} \Gamma_{t|t} \\ O_t \end{pmatrix} \\
\nu_{t|T} &= \nu_{T|T} \\
\Delta_{t|T} &= \begin{pmatrix} \Delta_{t|t} & 0 \\ 0 & \tilde{\Delta}_t \end{pmatrix}
\end{aligned}$$

with

$$\tilde{\Delta}_t = \begin{pmatrix} \Delta_\eta & 0 \\ 0 & \tilde{\Delta}_{t+1} \end{pmatrix} + \begin{pmatrix} \Gamma_\eta \\ O_{t+1} \end{pmatrix} (\Sigma_{t+1|T} - M_t \Sigma_{t|T} M_t') \begin{pmatrix} \Gamma_\eta \\ O_{t+1} \end{pmatrix}'.$$

References

Arellano-Valle, R. B., & Azzalini, A. (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis*, *99*, 1362–1382. doi:10.1016/j.jmva.2008.01.020.

Atkinson, T., Richter, A., & Throckmorton, N. (2019). The zero lower bound and estimation accuracy. *Journal of Monetary Economics*, . doi:10.1016/j.jmoneco.2019.06.007.

Blanchard, O. J., & Kahn, C. M. (1980). The Solution of Linear Difference Models under Rational Expectations. *Econometrica*, *48*, 1305–1311. doi:10.2307/1912186.

Diebold, F. X., Rudebusch, G. D., & Aruoba, B. S. (2006). The macroeconomy and the yield curve: A dynamic latent factor approach. *Journal of Econometrics*, *131*, 309–338. doi:10.1016/j.jeconom.2005.01.011.

Flecher, C., Naveau, P., & Allard, D. (2009). Estimating the closed skew-normal distribution parameters using weighted moments. *Statistics & Probability Letters*, *79*, 1977–1984. doi:10.1016/j.spl.2009.06.004.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications - A Journey Beyond Normality*. S.l.: CRC PRESS.

Guljanov, G., Mutschler, W., & Trede, M. (2022). *Pruned Skewed Kalman Filter and Smoother: With Applications to the Yield Curve*. CQE Working Papers 101 Center for Quantitative Economics (CQE), University of Münster. URL: https://www.wiwi.uni-muenster.de/cqe/sites/cqe/files/CQE_Paper/cqe_wp_101_2022.pdf.

Gupta, A. K., González-Farías, G., & Domínguez-Molina, J. (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis*, *89*, 181–190. doi:10.1016/S0047-259X(03)00131-3.

Harvey, A. C., & Phillips, G. D. A. (1979). Maximum Likelihood Estimation of Regression Models with Autoregressive-Moving Average Disturbances. *Biometrika*, *66*, 49. doi:10.2307/2335241.

Ireland, P. N. (2004). Technology Shocks in the New Keynesian Model. *Review of Economics and Statistics*, *86*, 923–936. doi:10.1162/0034653043125158.

Käärik, M., Selart, A., & Käärik, E. (2015). On Parametrization of Multivariate Skew-Normal Distribution. *Communications in Statistics - Theory and Methods*, *44*, 1869–1885. doi:10.1080/03610926.2012.760277.

Rezaie, J., & Eidsvik, J. (2014). Kalman filter variants in the closed skew normal setting. *Computational Statistics & Data Analysis*, *75*, 1–14. doi:10.1016/j.csda.2014.01.014.