# Introduction to Robust Kalman Filters

Janoś Gabler

December 6, 2018

# 1 Background

The goal is to estimate a dynamic non-linear latent factor model that can be expressed in the following form:

$$\mathbf{x}_{t+1} = F_t(\mathbf{x}_t) + \eta_t \qquad TransitionEquations \tag{1}$$

where $\mathbf{x}_t$ is an unobserved state vector of length $N$ in period $t$, $F_t(\mathbf{x}_t)$ is a parametric function with unknown parameters and $\eta_t$ is a vector of shocks. Let $\mathbf{Q}_t$ denote the covariance matrix of $\eta_t$.

$$\mathbf{y}_t = \mathbf{H_t}\mathbf{x}_t + \epsilon_t \qquad MeasurementEquations \tag{2}$$

where $\mathbf{y}_t$ is a vector of observable measurements with length $L_t$, $\mathbf{H}_t$ is a matrix of factor loadings and $\epsilon_t$ is a vector of measurement errors. Let $\mathbf{R}_t$ be the covariance matrix of $\epsilon_t$.

The following material is based on Cunha, Heckman, and Schennach, 2010 (CHS).

# 2 Efficient State Estimation

## 2.1 Preliminaries

Assume for a moment that the transition function $F_t$ (including parameters) as well as the matrices $\mathbf{H_t}$, $\mathbf{Q}_t$ and $\mathbf{R}_t$ are known for all $t \in T$ but the state vectors $\mathbf{x}_t$ are unknown and have to be estimated from measurements $\mathbf{y}_t$. This problem is known as optimal state estimation, which is an important and well researched topic in physics and engineering.

To efficiently estimate the state vector in period $t$, an estimator should not only use measurements from this period, but also take the information from all previous measurements into account. For linear systems, Kalman filters are the method of choice for state estimation (Kalman, 1960). For nonlinear systems, several nonlinear variants of the Kalman filter have been developed. Kalman filters treat the state of a system itself as random vector. Therefore, they are sometimes classified as Bayesian filters.

Kalman filters consist of a predict and an update step. They are initialised with an initial estimate for the mean $\bar{\mathbf{x}}_0$ and covariance matrix $\mathbf{P}_0$ of the distribution of the state vector. Then, in each period, the new measurements are incorporated to update the mean and covariance matrix of the state vector. After that, the transition equation is used to predict the mean and covariance matrix of the state vector in the next period. This predicted state vector can then again be updated with measurements.

So far, the only assumptions that have been made about the error terms are that they are additively separable and their means are zero. For the application of Kalman filters, the following assumptions will be made:

1. $\eta_t \sim \mathcal{N}(\mathbf{0_N}, \mathbf{Q_t})$ where $\mathbf{0}_N$ denotes a vector of zeros of length $N$, $\mathbf{Q_t}$ is a diagonal matrix.
2. The $\eta_t^j$ are serially independent over all $t$.
3. $\epsilon_t \sim \mathcal{N}(\mathbf{0}_{L_t}, \mathbf{R}_t)$ where $\mathbf{R}_t$ is a diagonal matrix.
4. The $\epsilon_{t,l}$ are serially independent over all $t$.
5. $\epsilon_{t,l}$ and $\eta_t^j$ are independent of $\mathbf{x}_t$ for all $t = 1, \dots, T$, $l = 1, \dots, L$ and each factor $j$.
6. The conditional density of the state vector $p(\mathbf{x}_t | y_{l,l})$ can be accurately approximated by a normal density for all $t = 1, \dots, T$ and all $l = 1, \dots, L_t$.[1]

Due to the assumption of a linear measurement system, the state vector can be estimated by combining the update step of a linear Kalman filter with the predict step of a nonlinear Kalman filter.[2] Apart from that, it will be convenient not to incorporate all measurements at once but to perform a separate update step for each measurement.

## 2.2 The Update Step of the Kalman Filter

The aim of the Kalman update is to efficiently combine information from measurements in the current period with previous measurements. To do so, the measurement function is used to convert the pre-update state vector into predicted measurements for the current period (equation 3). The difference between the predicted and actual measurements is called residual (equation 4). This residual, scaled by the so called Kalman gain, is then added to the pre-update state vector (equation 8). The Kalman gain is smaller if the variance of the measurement (calculated by equation 6) is large. This has the intuitive consequence that noisy measurements receive a low weight. The Kalman gain becomes larger if the pre-update covariance matrix has large diagonal entries (equation 5 and 7). Thus, measurements receive more weight if the pre-update state is known imprecisely due to bad initial values or a high process noise, for example. After the incorporation of the measurements, the state is always known with the same or more precision than before. This is reflected by subtracting a matrix with nonnegative diagonal elements from the pre-update covariance matrix (equation 9).

Let $\bar{\mathbf{x}}_{t|y_{t,l}^-}$ denote the mean of the conditional distribution of the state vector given all measurements up to but not including the $l^{th}$ measurement in period $t$. Let $\mathbf{P}_{t|y_{t,l}^-}$ denote the covariance matrix of this distribution. Let $\mathbf{h}_{t,l}$ denote the $l^{th}$ row of $\mathbf{H}_t$. Let $r_{t,l,l}$ be

[1] This assumption can be relaxed to the assumption that the density of the state vector can be approximated by a mixture of normal distributions. This relaxation can be found in appendix 6 of Cunha, Heckman, and Schennach, 2010.

[2] For the estimator with a nonlinear measurement system see appendix 6 of Cunha, Heckman, and Schennach, 2010

the $l^{th}$ diagonal element of $\mathbf{R}_t$. The update step that incorporates the $l^{th}$ measurement into the estimate is given by the following equations:

(3) $\bar{y}_{t,l|y_{t,l}^-} = \mathbf{h}_{t,l}\bar{\mathbf{x}}_{t|y_{t,l}^-}$ $\quad\quad$ $\bar{y}_{t,l|y_{t,l}^-} = E(y_{t,l}|y_{t,l}^-)$

(4) $\delta_{t,l} = y_{t,l} - \bar{y}_{t,l|y_{t,l}^-}$ $\quad\quad$ $\delta_{t,l}$ can be interpreted as residual

(5) $\mathbf{f}_{t,l} = \mathbf{P}_{t|y_{t,l}^-}\mathbf{h}_{t,l}^T$ $\quad\quad$ $\mathbf{f}_{t,l}$ is an intermediate result

(6) $\sigma_{t,l} = \mathbf{h}_{t,l}\mathbf{f}_{t,l} + r_{t,l,l}$ $\quad\quad$ $\sigma_{t,l}$ is the variance of $y_{t,l}$

(7) $\mathbf{k}_{t,l} = \frac{1}{\sigma_{t,l}}\mathbf{f}_{t,l}$ $\quad\quad$ $\mathbf{k}_{t,l}$ is the (scaled) Kalman gain

(8) $\bar{\mathbf{x}}_{t|y_{t,l}} = \bar{\mathbf{x}}_{t|y_{t,l}^-} + \mathbf{k}_{t,l}\delta_{t,l}$ $\quad\quad$ $\bar{\mathbf{x}}_{t|y_{t,l}}$ is the updated mean

(9) $\mathbf{P}_{t|y_{t,l}} = \mathbf{P}_{t|y_{t,l}^-} - \frac{1}{\sigma_{t,l}}\mathbf{f}_{t,l}\mathbf{f}_{t,l}^T$ $\quad\quad$ $P_{t|y_{t,l}}$ is the updated covariance matrix

## 2.3 The Predict Step of the Kalman Filter

In linear systems, the mean and covariance matrix of the system can be propagated to the next period by simply applying the linear transition equation. With a nonlinear transition function, however, this is not possible, as $E(f(X) \neq f(E(X)$ for a general nonlinear function $f$. For the nonlinear predict step, two basic options exist: The *extended Kalman filter* and the *unscented Kalman filter*. CHS choose the unscented Kalman filter because it has been shown to be more reliable in a wide range of settings (Van Der Merwe, 2004).

The intuition of the predict step of the unscented Kalman filter is relatively simple: firstly, a deterministic sample of points in the state space, called sigma points (equation 10), and accompanying weights are chosen (equation 11). Usually these are $2N + 1$ points and weights, where $N$ is the length of the state vector. Secondly, these sigma points are transformed using the true nonlinear transition equation. Thirdly, the weighted sample mean is used as estimate for the next period mean of the state vector (equation 12). Fourthly, the sum of the covariance matrix of the process noise and the weighted sample covariance of the transformed sigma points is used as estimate of the covariance matrix of the state vector (equation 13). Intuitively, the addition of the process noise accounts for the fact that the prediction always adds some uncertainty about the state of the system.

For the choice of sigma points and sigma weights, many different algorithms exist. All have in common that some form of matrix square root of the covariance matrix of the state vector is taken. Two definitions of matrix square root exist: 1) $\mathbf{A}$ is a matrix square root of $\mathbf{P}$ if $\mathbf{P} = \mathbf{AA}$. 2) $\mathbf{A}$ is a matrix square root of $\mathbf{P}$ if $\mathbf{P} = \mathbf{AA}^T$. The matrix square root is not unique in general and some matrices do not have a square root. However, all symmetric positive semi-definite matrices, i.e. all valid covariance matrices, can be decomposed into

$\mathbf{P} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{L}$ is lower triangular (Zhang, 1999). For the unscented Kalman filter, both definitions of matrix square root work. Below, the sigma point algorithm proposed by Julier and Uhlmann, 1997, is presented without reference to a particular type of matrix square root:

Let $\kappa \in \mathbb{R}$ be a scaling parameter. Usually, $\kappa$ is set to 2 if the distribution of the state vector is assumed to be normal. Let $\mathbf{P}_{t|t}$ denote the covariance matrix of the state vector, conditional on all measurements up to and including period $t$. Define $\mathbf{S}_{t|t} \equiv \sqrt{\mathbf{P}_{t|t}}$ as the matrix square root of $\mathbf{P}_{t|t}$ and let $\mathbf{s}_{t,n}$ denote its $n^{th}$ row.

Sigma points are calculated according to the following equations:

$$
\begin{aligned}
\chi_{t,n} &= \bar{\mathbf{x}}_{t|t} && \text{for } n = 0 \\
\chi_{t,n} &= \bar{\mathbf{x}}_{t|t} + \sqrt{N+\kappa}\ \mathbf{s}_{t,n} && \text{for } n = 1, \dots N \\
\chi_{t,n} &= \bar{\mathbf{x}}_{t|t} - \sqrt{N+\kappa}\ \mathbf{s}_{t,n} && \text{for } n = N+1, \dots 2N
\end{aligned} \tag{10}
$$

where $\chi_{t,n}$ is the $n^{th}$ sigma point at period $t$ that is calculated after incorporating all measurements of that period. The corresponding sigma weights are calculated as follows:

$$
\begin{aligned}
w_{t,n} &= \frac{\kappa}{N+\kappa} && \text{for } n = 0 \\
w_{t,n} &= \frac{1}{2(N+\kappa)} && \text{for } n = 1, \dots, 2N
\end{aligned} \tag{11}
$$

where $w_{t,n}$ is the $n^{th}$ sigma weight. Define $\widetilde{\chi}_{t,n} \equiv F_t(\chi_{t,n})$ where $F_t(\cdot)$ is defined as in equation 1. Then the predict step of the unscented Kalman filter is given by:

$$
\bar{\mathbf{x}}_{t+1|t} = \sum_{n=0}^{2N} w_{t,n}\widetilde{\chi}_{t,n} \tag{12}
$$

$$
\mathbf{P}_{t+1|t} = \left[ \sum_{n=0}^{2N} w_{t,n}(\widetilde{\chi}_{t,n} - \bar{\mathbf{x}}_{t+1|t})(\widetilde{\chi}_{t,n} - \bar{\mathbf{x}}_{t+1|t})^T \right] + \mathbf{Q_t} \tag{13}
$$

# 3 The Likelihood Interpretation of the Kalman Filter

Of course, the parameters of the function $F_t$ and the matrices $\mathbf{H}_t$, $\mathbf{Q}_t$ and $\mathbf{R}_t$ are unknown in reality. However, they can be estimated by maximum likelihood. The direct maximization of the likelihood function would involve the evaluation of high dimensional integrals which is computationally very expensive (Cunha, Heckman, and Schennach, 2010). Instead, CHS propose to use the Kalman filter approach described in the previous section to

reduce the number of computations required for each evaluation of the likelihood function dramatically.

To see how, define $\theta$ as the vector with all estimated parameters of the model. Then, the likelihood contribution of individual $i$ is given by:

$$\mathcal{L}(\theta|\mathbf{y}_1, \dots, \mathbf{y}_T) \equiv p_\theta(\mathbf{y}_1, \dots, \mathbf{y}_T) = \prod_{t=1}^{T} \prod_{l=1}^{L_t} p_\theta(y_{t,l}|y_{t,l}^-) \tag{14}$$

where $p_\theta(\mathbf{y}_1, \dots, \mathbf{y}_T)$ denotes the joint density of all measurements for individual $i$, conditional on the parameter vector $\theta$ and $p_\theta(y_{t,l}|y_{t,l}^-)$ is the density of the $l^{th}$ measurement in period $t$, given all measurements up to but not including this measurement. The subscript $i$ is again omitted for readability.

To see how this relates to the Kalman filter, recall that for each $t = 1, \dots, T$ and each $l = 1, \dots, L_t$, equation 3 calculates $\bar{y}_{t,l|y_{t,l}^-}$, i.e the expected value of the $l^{th}$ measurement in period $t$, conditional on all previous measurements. In addition, due to the normality and independence assumptions on the error terms and the factor distribution, $y_{t,l}$ is normally distributed around $\bar{y}_{t,l|y_{t,l}^-}$. Equation 6 can be used to calculate the variance $\sigma_{t,l}$ of this distribution. Thus, $p_\theta(y_{t,l}|y_{t,l}^-) = \phi_{\bar{y}_{t,l|y_{t,l}^-},\sigma_{t,l}}(y_{t,l})$ where $\phi_{\mu,\sigma}(\cdot)$ is the density of a normal random variable with mean $\mu$ and variance $\sigma$.

A nice feature of the estimator based on this factorization of the likelihood function is that it can deal very well with missing observations. If measurement $y_{t,l}$ is missing for inidividual $i$, the corresponding update of the state vector is just skipped. Technically, this means that the missing measurement is integrated out from the likelihood function.

# 4 Making the Estimator More Robust

## 4.1 Discussion of Numerical Instabilities

While the Kalman filter based maximum likelihood estimator proposed by CHS is statistically and computationally efficient, it is numerically unstable. The numerical instability caused by computer roundoff error is inherent to Kalman filters and has been discovered soon after Kalman published his original article. Since then, the precision of computers has increased enormously such that nowadays numerical problems are not a big issue for well specified Kalman filters. However, during the maximization of the likelihood function the maximization algorithm might pick parameter combinations that are far from leading to a well specified filter.

The numerical problems manifest themselves in two places:

1. In the update step, the subtraction in equation 9 can lead to negative diagonal elements in the updated covariance matrix of the state vector. While this is mathemati-

cally impossible in a well specified Kalman filter, numerical imprecisions and badly specified Kalman filters during the maximization process make it possible.

2. Even if the covariance matrix of the state vector has nonnegative diagonal entries, numerical imprecisions might render it not positive semi-definite. With this the existence of a matrix square root is not guaranteed, which can make the calculation of sigma points impossible.

CHS are aware of these problems. To solve the first problem, they recommend to find good initial values for the maximization by first constraining some parameters and letting the code find good initial values for the others. For the second problem, they propose to set all off-diagonal elements of $\mathbf{P}$ to zero before taking the square root, which then corresponds to taking the element wise square root of the diagonal elements. While this makes the estimator more robust, it is not standard practice in Kalman filtering and it is not guaranteed that an estimator based on this type of matrix square root produces reliable results.

## 4.2 Outline of the Solution

A better approach is to use a square root implementation of the Kalman filter. Many different square root Kalman filters exist. They are mathematically equivalent to normal Kalman filters but numerically more stable.

Instead of propagating the full covariance matrix of the state vector, square root Kalman filters propagate the square root of this matrix. This has three advantages:

1. It avoids overflow errors due to numbers with very small or large absolute values, as taking the square root makes large numbers smaller and small numbers larger.
2. By using a matrix square root $\mathbf{A}$ of the type $\mathbf{P} = \mathbf{A}\mathbf{A}^T$, the problematic covariance matrix is guaranteed to be positive semi-definite (Zhang, 1999), i.e. a valid covariance matrix. In particular, its diagonal entries are sums of squared terms and, consequently, guaranteed to be nonnegative. This solves the first problem.
3. By choosing an appropriate pair of square root update and predict algorithms, taking matrix square roots can be completely avoided. This eliminates the second problem.

The computational requirements of square root filters are comparable to those of normal Kalman filters. In the nonlinear case, they are even lower. For a maximally robust estimator, we use a pair of square root update and predict algorithms that completely avoid taking matrix square roots. The algorithm for the update was developed by Prvan and Osborne (Prvan and Osborne, 1988). The unscented square root predict step was proposed by Van Der Merwe and Wan (Van Der Merwe and Wan, 2001). Both propagate the transpose of a lower triangular matrix square root of the state covariance matrix.

## 4.3 The QR Decomposition of a Matrix

Both square root algorithms rely on a matrix factorization called QR decomposition. Note that in this subsection, **Q** and **R** do not denote the covariance matrices of the process and measurement noise but factors into which a matrix is decomposed.

**QR** is called QR decomposition of an $m \times n$ matrix **A** with $m \geq n$ if:

1. $\mathbf{A} = \mathbf{QR}$
2. **Q** is an orthogonal $m \times m$ matrix
3. **R** is an $m \times n$ matrix and the first $n$ rows of **R** form a upper triangular matrix and its remaining rows only contain zeros

The QR decomposition of a matrix always exists but is not unique. A useful property of the QR decomposition is that:

$$\mathbf{A}^T\mathbf{A} = (\mathbf{QR})^T\mathbf{QR} = \mathbf{R}^T\mathbf{Q}^T\mathbf{QR} = \mathbf{R}^T\mathbf{R} \tag{15}$$

where the last equality comes from the defining property of orthogonal matrices that $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$, where **I** denotes the identity matrix. Thus, the upper triangular part of **R** is the transpose of a lower triangular matrix square root of $\mathbf{A}^T\mathbf{A}$. For convenience, let $qr(\mathbf{A})$ denote the QR decomposition of **A** that only returns the upper triangular part of the matrix **R**.

## 4.4 The Update Step of the Square-Root Kalman Filter

Let $\mathbf{S}_{t|y_{l,t}^-}$ be a lower triangular matrix square root of $\mathbf{P}_{t|y_{l,t}^-}$ and keep the rest of the notation as in in section 2. Then, the square root update that incorporates the $l^{th}$ measurement in period $t$ is given by the following equations:

$\bar{y}_{t,l|y_{t,l}^-}$ and $\delta_{t,l}$ are calculated as in equation 3 and 4 respectively. Then the following intermediate results are calculated.

$$\mathbf{f}_{t,l}^* = \mathbf{S}_{t|y_{t,l}^-}^T \mathbf{h}_{t,l}^T \tag{16}$$

$$\mathbf{M}_{t,l} = \begin{bmatrix} \sqrt{r_{t,l,l}} & \mathbf{0}_N \\ \mathbf{f}_{t,l}^* & \mathbf{S}_{t|y_{t,l}^-}^T \end{bmatrix} \tag{17}$$

It can be shown that:

$$qr(\mathbf{M}_{t,l}) = \begin{bmatrix} \sqrt{\sigma_{t,l}} & \frac{1}{\sqrt{\sigma_{t,l}}}\mathbf{f}_{t,l}^T \\ \mathbf{0} & \mathbf{S}_{t|y_{t,l}}^T \end{bmatrix} \tag{18}$$

where $\mathbf{S}^T_{t|y_{t,l}}$ denotes the transpose of a lower triangular square root of the updated covariance matrix. The proof that equation 18 holds is not very insightful and can be found in the technical appendix.

The matrix in equation 18 also contains $\mathbf{f}_{t,l}$ and $\sigma_{t,l}$ such that the Kalman gain can be calculated as in equation 7 and the mean of the state vector can be updated as in equation 8, which completes the Kalman update.

## 4.5 The Predict Step of the Square-Root Kalman Filter

For the square root implementation of the unscented predict step in period $t$, firstly the sigma points are calculated as in equation 10, where this time $\mathbf{S}_{t|t}$ is required to be a lower triangular matrix square root of $\mathbf{P}_{t|t}$. Again, $\tilde{\mathcal{X}}_t$ denotes the $(2N+1)\times N$ matrix of the transformed sigma points. The calculation of the predicted mean of the state vector remains the same as before (equation 12).

Define $\mathbf{A}_t$ as stacked matrix of of weighted deviations of the sigma points from the predicted mean and the covariance matrix of the transition shocks:

$$\mathbf{A}_t \equiv \begin{bmatrix} \sqrt{w_{t,0}}(\tilde{\mathcal{X}}_{t,0} - \bar{\mathbf{x}}_{t+1|t}) \\ \cdots \\ \sqrt{w_{t,2n}}(\tilde{\mathcal{X}}_{t,2n} - \bar{\mathbf{x}}_{t+1|t}) \\ \sqrt{\mathbf{Q}_t} \end{bmatrix} \tag{19}$$

Then equation 13 can be rewritten as:

$$\mathbf{P}_{t+1|t} = \mathbf{A}_t^T \mathbf{A}_t \tag{20}$$

and by the relation of the QR decomposition and and the lower triangular matrix square root (equation 15) a lower triangular matrix square root of $\mathbf{P}_{t+1|t}$ is given by $qr(\mathbf{A}_t)^T$.

# References


Cunha, Flavio, James Heckman, and Susanne M. Schennach (2010). "Estimating the Technology of Cognitive and Noncognitive Skill Formation". In: *Econometrica* 78.3, pp. 883–931. ISSN: 1468-0262. DOI: 10.3982/ECTA6551. URL: http://dx.doi.org/10.3982/ECTA6551.

Julier, Simon J. and Jeffrey K. Uhlmann (1997). "New extension of the Kalman filter to nonlinear systems". In: *Proc. SPIE* 3068, Signal Processing, Sensor Fusion, and Target Recognition VI, pp. 182–193. DOI: 10.1117/12.280797. URL: http://dx.doi.org/10.1117/12.280797.

Kalman, R.E. (1960). "A New Approach to Linear Filtering and Prediction Problems". In: *ASME Transactions (Journal of Basic Engineering)* 82, Part D, pp. 35–45. DOI: 10.1115/1.3662552.

Prvan, Tania and M. R. Osborne (1988). "A Square-Root Fixed-Interval Discrete-Time Smoother". In: *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics* 30.1, pp. 57–68. DOI: /10.1017/S0334270000006032. URL: http://dx.doi.org/10.1017/S0334270000006032.

Van Der Merwe, R. and E.A. Wan (2001). "The square-root unscented Kalman filter for state and parameter-estimation". In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*. Vol. 6, 3461–3464 vol.6. DOI: 10.1109/ICASSP.2001.940586.

Van Der Merwe, Rudolph (2004). "Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models". PhD thesis. OGI School of Science & Engineering at Oregon Health & Science University. URL: http://www.cslu.ogi.edu/publications/ps/merwe04.pdf.

Zhang, Fuzhen (1999). *Matrix Theory: Basic Results and Techniques*. Universitext (Berlin. Print). Springer. ISBN: 9780387986968. URL: https://books.google.de/books?id=z2hOMmPISNoC.

# Technical Appendix

## Proof of Equation 18 In the Derivation of the Square-Root Update

To see why equation 18 in the derivation of the square root Kalman filter update is true, define $\mathbf{U}_{t,l} \equiv qr(\mathbf{M}_{t,l})$ and partition it as follows:

$$\mathbf{U}_{t,l} = \begin{bmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} \\ \mathbf{0} & \mathbf{U}_{2,2} \end{bmatrix} \tag{21}$$

where $\mathbf{U}_{1,1}$ is a scalar, $\mathbf{U}_{1,2}$ a row vector of length $N$, $\mathbf{0}$ a column vector of length $N$ filled with zeros and $\mathbf{U}_{2,2}$ an upper triangular $N \times N$ matrix. Recall from the definition of $\mathbf{U_{t,l}}$ and equation 15 that $\mathbf{U}_{t,l}^T \mathbf{U}_{t,l} = \mathbf{M}_{t,l}^T \mathbf{M}_{t,l}$. Multiplying out both sides of this equality yields:

$$\begin{bmatrix} r_{t,l,l} + \mathbf{f}_{t,l}^{*T} \mathbf{f}_{t,l}^{*} & \mathbf{f}_{t,l}^{*T} \mathbf{S}_{t|y_{t,l}^-}^T \\ \mathbf{S}_{t|y_{t,l}^-} \mathbf{f}_{t,l}^{*} & \mathbf{S}_{t|y_{t,l}^-} \mathbf{S}_{t|y_{t,l}^-}^T \end{bmatrix} = \begin{bmatrix} \mathbf{U}_{1,1}^2 & \mathbf{U}_{1,1}\mathbf{U}_{1,2} \\ \mathbf{U}_{1,2}^T \mathbf{U}_{1,1} & \mathbf{U}_{1,2}^T \mathbf{U}_{1,2} + \mathbf{U}_{2,2}^T \mathbf{U}_{2,2} \end{bmatrix} \tag{22}$$

It is obvious from equation 6 and 16 that $\mathbf{U}_{1,1} = \sqrt{\sigma_{t,l}}$. Using this and noting that $\mathbf{f}_{t,l}^{*T} \mathbf{S}_{t|y_{t,l}^-}^T = \mathbf{f}_{t,l}^T$, where $\mathbf{f}_{t,l}$ is defined as in equation 5, one obtains that:

$$\mathbf{U}_{1,2} = \frac{\mathbf{f}_{t,l}^T}{\sqrt{\sigma_{t,l}}} \tag{23}$$

It remains to show that $\mathbf{U}_{2,2} = \mathbf{S}_{t|y_{t,l}}^T$. By noting that the the bottom right element of the left hand side of equation 22 is, by definition, equal to the pre-update covariance matrix $\mathbf{P}_{t|y_{t,l}^-}$ and plugging in the value for $\mathbf{U}_{1,2}$, one obtains that:

$$\mathbf{U}_{2,2}^T \mathbf{U}_{2,2} = \mathbf{P}_{t|y_{t,l}^-} - \frac{1}{\sigma_{t,l}} \mathbf{f}_{t,l} \mathbf{f}_{t,l}^T = \mathbf{P}_{t|y_{t,l}} \tag{24}$$

where the last equality comes comes from equation 9. Thus $\mathbf{U}_{2,2}^T$ is a matrix square root of $\mathbf{P}_{t|y_{t,l}}$ and by the definition of the QR decomposition it is lower triangular, which completes the proof.