1 Basic Linear State Space

1.1 Definition

$$x_t = Fx_{t-1} + \epsilon_t^x \tag{1}$$
$$y_t = Hx_t + \epsilon_t^y \tag{2}$$

$x_t \in \mathbb{R}^{d_x}$ and $y_t \in \mathbb{R}^{d_y}$.

$$\epsilon_t^x \sim \mathcal{N}(0, Q) \tag{3}$$
$$\epsilon_t^y \sim \mathcal{N}(0, R) \tag{4}$$

Q and R are positive definite covariance matrices.

Equivalently in terms of transition and observation densities:

$$p(x_t|x_{t-1}, F, Q) = \mathcal{N}(x_t|Fx_{t-1}, Q) \tag{5}$$
$$p(y_t|x_t, H, R) = \mathcal{N}(y_t|Hx_t, R) \tag{6}$$

The initial state x_1 may be known or may be assigned a Gaussian prior.

1.2 Identifiability

1.3 Bayesian Learning with Gibbs Sampling

Target the joint distribution $p(F, Q, x_{1:T}|y_{1:T}$. Gibbs sampling can be used, targeting the state and parameter posterior conditionals alternately. State conditional sampled by Kalman filtering and backward simulation.

1.3.1 Parameter Conditional

The conjugate prior is:

$$p(F, Q) = p(F|Q)p(Q) \tag{7}$$

$$Q \sim \mathcal{IW}(\nu_0, \Psi_0) \tag{8}$$
$$F|Q \sim \mathcal{MN}(M_0, Q, \Omega_0) \tag{9}$$

So the prior densities are:

$$p(Q) = \frac{|\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0}{2}}\Gamma\left(\frac{\nu_0}{2}\right)} |Q|^{-\frac{\nu_0+d_x+1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}\Psi_0\right]\right) \tag{10}$$
$$p(F|Q) = |2\pi Q|^{-\frac{1}{2}} |2\pi\Omega_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}(F-M_0)\Omega_0^{-1}(F-M_0)^T\right]\right) \tag{11}$$

$\nu_0 \in \mathbb{R}, \nu_0 > d_x$. Ψ_0 and Ω_0 are $d_x \times d_x$ positive definite matrices. $M_0 \in \mathbb{R}^{d_x \times d_x}$.

The likelihood function is:

$$\begin{aligned} p(x_{1:T}|F, Q) &= p(x_0) \prod_{t=2}^{T} p(x_t|x_{t-1}, F, Q) \\ &\propto |Q|^{-\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\sum_{t=2}^{T}(x_t - Fx_{t-1})^T Q^{-1}(x_t - Fx_{t-1})\right) \\ &= |Q|^{-\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}\sum_{t=2}^{T}(x_t - Fx_{t-1})(x_t - Fx_{t-1})^T\right]\right) \\ &= |Q|^{-\frac{1}{2}S_0} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}\left(FS_1F^T - FS_2^T - S_2F^T + S_3\right)\right]\right) \end{aligned} \tag{12}$$

where the sufficient statistics are:

$$S_0 = T - 1 \tag{13}$$

$$S_1 = \sum_{t=2}^{T} x_{t-1} x_{t-1}^T \tag{14}$$

$$S_2 = \sum_{t=2}^{T} x_t x_{t-1}^T \tag{15}$$

$$S_3 = \sum_{t=2}^{T} x_t x_t^T \tag{16}$$

$$\tag{17}$$

The parameter posterior conditional is:

$$p(F, Q | x_{1:T}, y_{1:T}) = p(F|Q, x_{1:T}) p(Q|x_{1:T}) \tag{18}$$

$$Q|x_{1:T} \sim \mathcal{IW}(\nu, \Psi) \tag{19}$$

$$F|Q, x_{1:T} \sim \mathcal{MN}(M, Q, \Omega) \tag{20}$$

Hyperparameter updates:

$$\Omega^{-1} = \Omega_0^{-1} + S_1 \tag{21}$$

$$M\Omega^{-1} = M_0\Omega_0^{-1} + S_2 \tag{22}$$

$$\nu = \nu_0 + S_0 \tag{23}$$

$$\Psi = \Psi_0 + S_3 + M_0\Omega_0^{-1}M_0^T - M\Omega^{-1}M^T \tag{24}$$

Proof:

$$\begin{aligned}
p(F, Q | x_{1:T}, y_{1:T}) &\propto p(x_{1:T}|F, Q) p(F|Q) p(Q) \\
&= |Q|^{-\frac{1}{2}S_0} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}\left(FS_1F^T - FS_2^T - S_2F^T + S_3\right)\right]\right) \\
&\quad \times |Q|^{-\frac{\nu_0 + d_x + 1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}\Psi_0\right]\right) \\
&\quad \times |Q|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}(F - M_0)\Omega_0^{-1}(F - M_0)^T\right]\right) \\
&= |Q|^{-\frac{(\nu_0 + S_0) + d_x + 1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[Q^{-1}(\Psi_0 + S_3)\right]\right) \\
&\quad \times |Q|^{-\frac{1}{2}} \exp\Bigg(-\frac{1}{2}\operatorname{Tr}\Bigg[Q^{-1}\Big(F(\Omega_0^{-1} + S_1)F^T - F(\Omega_0^{-1}M_0^T + S_2^T) \\
&\qquad\qquad - (M_0\Omega_0^{-1} + S_2)F^T + M_0\Omega_0^{-1}M_0^T\Big)\Bigg]\Bigg)
\end{aligned} \tag{25}$$

Finish by completing the square and comparing terms.

2 Degenerate Linear State Space Model

2.1 Definition and Complications

$$x_t = Fx_{t-1} + G\epsilon_t^x \tag{26}$$

$$\epsilon_t^x \sim \mathcal{N}(0, I) \tag{27}$$

This is equivalent to the basic case, except that $Q = GG^T$ is only positive semi-definite. This is not in itself a problem. The probability distribution associated with each transition is still well-defined, although it does not have a density on $\mathbb{R}^{d_x}$. The state perturbations $(x_t - Fx_{t-1})$ must all lie in a linear subspace.

There are two problems with extending the Gibbs sampling methodology to such a model. First, the state sequence and the transition covariance matrix are too strongly related. The sampled state sequence defines the subspace, so we cannot freely change Q. Second, the conjugate prior is tricky. The matrix-normal-inverse-wishart can be generalised to the case where Q is not full rank (using singular wishart and singular matrix normal distributions). However, this is not helpful, because it constrains the transition matrix to lie in a linear subspace.

Because Wishart distributions are somewhat more studied than their inverse counterparts, it will be easier in the following section to work with the precision, defined as $\Upsilon = (GG^T)^+$. We will need the eigen-decomposition of this frequently.

$$\Upsilon = V \Lambda V^T \tag{28}$$

This is the "non-singular" factorisation, where Λ is a diagonal $r \times r$ matrix with $r = \text{Rank}(\Upsilon)$, and V from the appropriate Stiefel manifold. The full decomposition will also be useful at times.

$$\Upsilon = \begin{bmatrix} V & V_\perp \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V^T \\ V_\perp^T \end{bmatrix} \tag{29}$$

2.2 Conjugate Prior

2.2.1 Wishart Part

There are two ways to extend the Wishart distribution to positive semi-definite matrices. Either let $\nu_0 \in \{z \in \mathbb{N}^+ | z < d_x\}$, in which case $\text{Rank}(\Upsilon) = \nu_0$, or let Ψ_0 be positive semi-definite. The latter makes it tricky to work out what we should set Ψ_0 to, so we'll go with the former. If we condition on a known value of the rank r, then the singular Wishart density is,

$$p(\Upsilon) = \frac{\pi^{\frac{1}{2}r(d_x - r)} \left|\Psi_0\right|^{\frac{r}{2}}}{2^{\frac{1}{2}rd_x} \Gamma_r\left(\frac{r}{2}\right)} \left|\Lambda\right|^{\frac{1}{2}(r - d_x - 1)} \exp\left(-\frac{1}{2}\Upsilon\Psi_0\right). \tag{30}$$

[1, 5]

Note that Ψ_0 is defined as the inverse of what it usually is, for consistency with the basic model case.

2.2.2 Matrix Normal Part

It's fine for the row-variance matrix in a matrix normal distribution to be less than full rank. That just means that the matrix is constrained to lie in a particular subspace. There's even a density associated with this. See [1]. But this is no good as a prior. We want to be able to change the transition matrix freely.

Instead we can use,

$$F|\Upsilon \sim \mathcal{MN}\left(M_0, \Upsilon^+ + V_\perp \Lambda_\perp^{-1} V_\perp^T, \Omega_0\right) \tag{31}$$

$\Lambda_\perp$ is a diagonal matrix of positive eigenvalues which relax the constraint to lie in the subspace. They control the rate at which the density decays as we move away from that subspace. Setting these extra eigenvalues might be tricky. I haven't really worked out how to do this well yet. But they can depend on the other ones, so we could take the minimum, maximum or average of the others. Making them really big is uninformative.

2.3 MCMC for Bayesian Learning

A sampled state trajectory $x_{1:T}$ defines the subspace in which the state perturbations must lie, which means that neither F nor Υ can be freely changed, and Gibbs sampling alone does not work. However, we can still sample within the space specified by this constraint. Additional MCMC moves will then be needed which allow the subspace to be changed.

2.3.1 Factorising the Precision

Using Givens rotations, we can efficiently factorise the precision eigenvector matrix as follows.

$$\begin{bmatrix} V & V_\perp \end{bmatrix} = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & E_\perp \end{bmatrix} \begin{bmatrix} U_R & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & U_N \end{bmatrix} \tag{32}$$

So

$$V = UEU_R \tag{33}$$
$$V_\perp = U_\perp E_\perp U_N \tag{34}$$

Then defined the following, which is positive definite

$$D = EU_R \Lambda U_R^T E^T \tag{35}$$

So

$$\Upsilon = UDU^T \tag{36}$$

2.3.2 State Trajectory Likelihood

There is no proper transition density over $\mathbb{R}^{d_x}$ for the degenerate model. However, there is a density associated with the underlying state disturbances which allows us to write a likelihood function.

$$x_t = Fx_{t-1} + UD^{-\frac{1}{2}}\epsilon_t^x \tag{37}$$
$$\tag{38}$$

$$\Rightarrow U^T(x_t - Fx_{t-1}) \sim \mathcal{N}\left(0, D^{-1}\right) \tag{39}$$

$$p(x_t|x_{t-1}, F, Q) \propto |D|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x_t - Fx_{t-1})^T UDU^T (x_t - Fx_{t-1})\right) \tag{40}$$

The likelihood function associated with the state trajectory is:

$$\begin{aligned} p(x_{1:T}|F, Q) &= p(x_0) \prod_{t=2}^{T} p(x_t|x_{t-1}, F, Q) \\ &\propto |D|^{\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\sum_{t=2}^{T}(x_t - Fx_{t-1})^T UDU^T (x_t - Fx_{t-1})\right) \\ &= |D|^{\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[DU^T \sum_{t=2}^{T}(x_t - Fx_{t-1})(x_t - Fx_{t-1})^T U\right]\right) \\ &= |D|^{\frac{1}{2}S_0} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[D\left(U^T F S_1 F^T U - U^T F S_2^T U - U^T S_2 F^T U + U^T S_3 U\right)\right]\right) \quad (41) \\ &= |D|^{\frac{1}{2}S_0} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[D\left(F_U S_1 F_U^T - F_U S_2^T U - U^T S_2 F_U^T + U^T S_3 U\right)\right]\right) \quad (42) \end{aligned}$$

where

$$F_U = U^T F \tag{43}$$

This is all subject to the constraint that each state must be reachable from the previous one, which we can write concisely as,

$$X_{2:T} = FX_{1:T-1} + UZ \tag{44}$$

where $Z \in \mathbb{R}^{r \times d_x}$ and

$$X_{2:T} = \begin{bmatrix} x_2 & x_3 & \dots & x_T \end{bmatrix} \tag{45}$$

2.3.3 Transforming the Prior

If we transform from (F, Υ) to (F_U, D), then the prior on these variables is as follows.

Using the transformation property of the matrix normal distribution:

$$F_U|\Upsilon = F_U|U, D \sim \mathcal{MN}\left(U^T M_0, D^{-1}, \Omega_0\right) \tag{46}$$

Using the transformation property of the Wishart distribution:

$$D|U \sim \mathcal{W}\left(r, U^T \Psi_0 U\right) \tag{47}$$

2.3.4 Within-Subspace Parameter Conditional

We can now get to the conditional distribution for $p(F_U, D|U, x_{1:T})$ which is well-behaved and can be sampled.

The parameter posterior conditional is:

$$p(F_U, D|U, x_{1:T}, y_{1:T}) = p(F_U|D, U, x_{1:T})p(D|U, x_{1:T}) \tag{48}$$

$$D|U, x_{1:T} \sim \mathcal{W}\left(\nu, \Psi\right) \tag{49}$$

$$F_U|D, U, x_{1:T} \sim \mathcal{MN}\left(M, Q, \Omega\right) \tag{50}$$

Hyperparameter updates:

$$\Omega^{-1} = \Omega_0^{-1} + S_1 \tag{51}$$

$$M\Omega^{-1} = U^T\left(M_0\Omega_0^{-1} + S_2\right) \tag{52}$$

$$\nu = r + S_0 \tag{53}$$

$$\Psi = U^T\left(\Psi_0 + S_3 + M_0\Omega_0^{-1}M_0^T\right)U - M\Omega^{-1}M^T \tag{54}$$

Proof:

$$\begin{aligned}
p(F_U, D|U, x_{1:T}) &\propto p(x_{1:T}|F_U, D, U)p(F_U|D, U)p(D|U) \\
&\propto |D|^{\frac{1}{2}S_0} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[D\left(F_U S_1 F_U^T - F_U S_2^T U - U^T S_2 F_U^T + U^T S_3 U\right)\right]\right) \\
&\quad \times |D|^{\frac{r-r-1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[DU^T\Psi_0 U\right]\right) \\
&\quad \times |D|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[D(F_U - U^T M_0)\Omega_0^{-1}(F_U - U^T M_0)^T\right]\right) \\
&= |D|^{\frac{(r+S_0)-r-1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left[DU^T(\Psi_0 + S_3)U\right]\right) \\
&\quad \times |D|^{\frac{1}{2}} \exp\Bigg(-\frac{1}{2}\operatorname{Tr}\Bigg[D\Big(F_U(\Omega_0^{-1} + S_1)F_U^T - F_U(\Omega_0^{-1}M_0^T + S_2^T)U \\
&\qquad\qquad - U^T(M_0\Omega_0^{-1} + S_2)F_U^T + U^T M_0\Omega_0^{-1}M_0^T U\Big)\Bigg]\Bigg)
\end{aligned} \tag{55}$$

Finish by completing the square and comparing terms.

Of course, we want F, not F_U. We can recover the matrix uniquely using the constraint from the likelihood. The previous value of F from before we sampled must have satisfied this constraint. Let's call that matrix F^*.

$$X_{2:T} = F^* X_{1:T-1} + UZ^*$$

$$X_{2:T} = FX_{1:T-1} + UZ$$

$$\Rightarrow 0 = (F - F^*)X_{1:T-1} + U(Z - Z^*) \tag{56}$$

$$\Rightarrow 0 = (F_U - F_U^*)X_{1:T-1} + (Z - Z^*) \tag{57}$$

$$\Rightarrow 0 = (F - F^*)X_{1:T-1} - U(F_U - F_U^*)X_{1:T-1} \tag{58}$$

$$0 = \left[(F - F^*) - U(F_U - F_U^*)\right]X_{1:T-1} \tag{59}$$

$$\Rightarrow 0 = (F - F^*) - U(F_U - F_U^*) \tag{60}$$

$$\begin{aligned} F &= F^* + U(F_U - F_U^*) \\ &= F^* + U(F_U - U^T F^*) \\ &= (I - UU^T)F^* + UF_U \end{aligned} \tag{61}$$

2.4 Metropolis-Hastings for the Precision Subspace

In order to change U, we can target $p(\Upsilon|F, y_{1:T})$ with Metropolis-Hastings, which constitutes a collapsed Gibbs move.

We sample a rotation matrix Ξ from some proposal distribution ς and then apply the following transformation,

$$\begin{bmatrix} \Upsilon' \\ \Xi' \end{bmatrix} = \begin{bmatrix} \Xi\Upsilon\Xi^T \\ \Xi^T \end{bmatrix},$$

which is clearly its own inverse. Since $|\Xi| = 1$, it is straightforward to show that the Jacobian of this transformation is 1, and hence using the reversible jump interpretation of Metropolis-Hastings [2,3], the acceptance probability is,

$$\begin{aligned} \alpha(\Upsilon \to \Upsilon') &= \min\left\{1, \frac{p(\Upsilon'|F, y_{1:T})\varsigma(\Xi')}{p(\Upsilon|F, y_{1:T})\varsigma(\Xi)}\right\} \\ &= \min\left\{1, \frac{p(y_{1:T}|F, \Upsilon')}{p(y_{1:T}|F, \Upsilon)} \times \frac{p(\Upsilon', F)}{p(\Upsilon, F)} \times \frac{\varsigma(\Xi')}{\varsigma(\Xi)}\right\}. \end{aligned} \tag{62}$$

The first term is simply a ratio of Kalman filter likelihoods.

There are numerous ways to sample the rotation matrix Ξ from a suitable proposal distribution ς. For example, we could use the Cayley transform [4], a bijective mapping from the skew-symmetric matrices to the rotation matrices, defined by,

$$P(S) = (I - S)^{-1}(I + S). \tag{63}$$

To sample from ς, we draw $\frac{1}{2}d_x(d_x - 1)$ independent scalar random variables $\{s_{i,j}\}_{0<i<j<d_x}$ from any zero-mean distribution; a nice choice is,

$$s_{k,l} \sim \mathcal{N}\left(0, \sigma_s^2\right). \tag{64}$$

Use these to construct a skew-symmetric matrix S,

$$S_{k,l} = \begin{cases} s_{k,l} & k < l \\ -s_{l,k} & k > l \\ 0 & k = l, \end{cases} \tag{65}$$

and then set $\Xi = P(S)$. The Cayley transform has the property that $P(-S) = P(S^T) = P(S)^{-1} = P(S)^T$, which implies that $\varsigma(\Xi) = \varsigma(\Xi^T) = \varsigma(\Xi')$, leading to a cancellation in the acceptance probability.

There is an alternative using Givens rotations. First sample $i \in [1, d_x]$, $j \in [1, d_x] \setminus i$, and $\gamma \in [-\pi/2, \pi/2]$ from some zero-mean distribution. Form the Givens matrix $\Gamma_{i,j}(\gamma)$ such that,

$$\left[\Gamma_{i,j}(\gamma) - I\right]_{k,l} = \begin{cases} \cos(\gamma) - 1 & k = l = i \text{ or } k = l = j \\ \sin(\gamma) & k = i, l = j \\ -\sin(\gamma) & k = j, l = i \\ 0 & \text{otherwise,} \end{cases} \tag{66}$$

and use $\Xi = \Gamma_{i,j}(\gamma)$. This also has the property that $\Gamma_{i,j}(-\gamma) = \Gamma_{i,j}(\gamma)^T$, meaning that we achieve the same cancellation of the proposals as before.

References

[1] José A Díaz-García, Ramón Gutierrez Jáimez, and Kanti V Mardia. Wishart and pseudo-wishart distributions and some applications to shape theory. *Journal of Multivariate Analysis*, 63(1):73 – 87, 1997.

[2] Peter J Green and David I Hastie. Reversible jump mcmc. *Genetics*, 155(3):1391–1403, 2009.

[3] P.J. Green. Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82(4):711, 1995.

[4] Carlos A. Len, Jean-Claude Mass, and Louis-Paul Rivest. A statistical model for random rotations. *Journal of Multivariate Analysis*, 97(2):412 – 430, 2006.

[5] Harald Uhlig. On singular wishart and singular multivariate beta distributions. *The Annals of Statistics*, 22(1):pp. 395–405, 1994.