Sparseness, conditional independence and the Kalman filter

Rasmus Waagepetersen

April 22, 2020

Outline:

1. conditional independence
2. sparseness and conditional independence for multivariate normal distributions
3. the Kalman filter (and smoother)

Sparseness and prediction - computational strategies

Consider prediction of U given Y. A computational challenge is to handle the inverse of $\mathbb{C}\mathrm{ov}\, Y = Z\Psi Z^{\mathsf{T}} + \Sigma$.

Using sparse matrix Cholesky as in miniproject is one solution.

Another solution: exploit conditional independence implied by sparseness $\Rightarrow$ Kalman filter.

Conditional independence

Suppose X, Y, Z are random variables (or vectors). Then we define X and Y to be conditionally independent given Z if

$$p(x, y|z) = p(x|z)p(y|z)$$

The following statements are equivalent:

1. $p(x, y|z) = p(x|z)p(y|z)$
2. $p(x, y, z) = f(x, z)g(y, z)$ for some functions f and g
3. $p(x|y, z) = p(x|z)$
4. $p(y|x, z) = p(y|z)$

($p(\cdot)$ generic notation for (possibly conditional) probability densities)

Suppose $X \sim N(\mu, \Sigma)$ with precision matrix $Q = \Sigma^{-1}$.

Then X_i and X_j conditionally independent given $X_{-\{i,j\}} \Leftrightarrow Q_{ij} = 0$. This follows from decomposition

$$(x-\mu)^{\mathsf{T}} Q (x-\mu) =$$
$$\left\{ (x_i - \mu_i)^2 Q_{ii} + 2 \sum_{k \neq i} (x_i - \mu_i)(x_k - \mu_k) Q_{ik} \right\} + \sum_{\substack{l,k: \\ l \neq i, k \neq i}} (x_l - \mu_l)(x_k - \mu_k) Q_{lk}$$

Note that x_i not in last term and $Q_{ij} = Q_{ji} = 0$ implies x_j not in first term. Thus we obtain factorization of density of X:

$$p(x) = f(x_i, x_{-\{i,j\}}) g(x_j, x_{-\{i,j\}})$$

In particular, if Q is sparse, a lot of X_i, X_j will be conditionally independent given the remaining variables.

Unnormalized density

To specify a probability density it is enough to specify an *unnormalized* density $h(\cdot)$:

$$f(x) \propto h(x) \Leftrightarrow f(x) = h(x)/c$$

where normalizing constant c uniquely determined by:

$$\int f(x)\mathrm{d}x = 1 \Leftrightarrow \int h(x)/c\mathrm{d}x = 1 \Leftrightarrow c = \int h(x)\mathrm{d}x$$

For example if X has density proportional to

$$h(x) = a^{x-1}\exp(-bx), \quad a, b > 0$$

we know that X has a Gamma distribution.

Conditional distribution of X_i

By previous slide

$$p(x_i|x_{-i}) \propto \exp(-\frac{1}{2}(x_i-\mu_i)^2 Q_{ii} - \sum_{k\neq i}(x_i-\mu_i)(x_k-\mu_k)Q_{ik})$$

Note for a normal distribution $Y \sim N(\xi, \sigma^2)$,

$$p(y) \propto \exp(-\frac{1}{2\sigma^2}y^2 + \frac{1}{\sigma^2}y\xi)$$

Comparing the two above equations we get

$$X_i|X_{-i} = x_{-i} \sim N(\mu_i - \frac{1}{Q_{ii}}\sum_{k\neq i} Q_{ik}(x_k-\mu_k), Q_{ii}^{-1})$$

Again we see that $Q_{ij} = Q_{ji} = 0 \Leftrightarrow X_i$ conditionally independent of X_j given $X_{-\{i,j\}}$.

Looking at bivariate distribution of (X_i, X_j) given $X_{-\{i,j\}}$ shows that the conditional (partial) correlation is

$$\mathbb{C}\mathrm{orr}[X_i, X_j|X_{-i,j}] = -Q_{ij}/\sqrt{Q_{ii}Q_{jj}}$$

A state-space model

Special case of linear mixed model:

$$U_1 \sim N(\mu_1, \Phi_1)$$
$$U_i = GU_{i-1} + W_i, W_i \sim N(0, \Phi)$$
$$Y_i = FU_i + V_i, V_i \sim N(0, \Sigma)$$

U_1, W_i, V_i all independent random vectors.

This is equivalent to $U_1 \sim N(\mu_1, \Phi_1)$ and for $i = 2, 3, \dots$

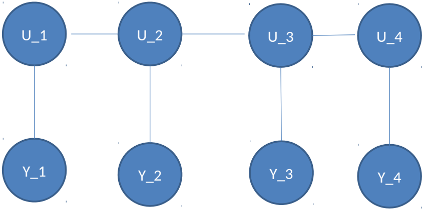
$$U_i | U_1 = u_1, \dots, U_{i-1} = u_{i-1}, Y_1 = y_1, \dots, Y_{i-1} = y_{i-1} \sim N(Gu_{i-1}, \Phi)$$
$$Y_i | U_1 = u_1, \dots, U_i = u_i, Y_1 = y_1, \dots, Y_{i-1} = y_{i-1} \sim N(Fu_i, \Sigma)$$

By factorization of joint density it follows that
$(U_1, \dots, U_{i-1}, Y_1, \dots, Y_{i-1})$, Y_i and $(U_{i+1}, \dots, U_n, Y_{i+1}, \dots, Y_n)$
are conditionally independent given U_i.

'past is independent of future given present state U_i'

Conditional independence graph

Conditional independences conveniently summarized by graph (edges correspond to equations defining model):



Two variables *not* joined by an edge iff they are conditionally independent given rest.

If two sets of variables are *separated* by a third set, then the two sets are independent given the third set.

The filtering problem

- is to predict U_n given $Y_{1:n}$.

I.e. we need to compute the normal distribution of U_n given $Y_{1:n}$.

The Kalman filter is a recursive algorithm for doing this.

We denote $\hat{u}_{n-1}$ and Σ_{n-1} the conditional mean and variance matrix of U_{n-1} given $y_{1:(n-1)}$. I.e.

$$U_{n-1}|Y_{1:(n-1)} = y_{1:(n-1)} \sim N(\hat{u}_{n-1}, \Sigma_{n-1})$$

(solution of filtering problem at 'time' $n-1$)

Useful observation I: 'a conditional density is just another probability density'

Consider X, Y given $Z = z$ with conditional densities $f(x|z)$, $f(y|z)$ and $f(x, y|z)$.

We can rename for a moment $g(x) = f(x|z)$, $g(y) = f(y|z)$, $g(x, y) = f(x, y|z)$. Then $g(x)$, $g(y)$ and $g(x, y)$ just like ordinary probability densities (but in the 'world' where $Z = z$)

In other words, if $(X, Y)|Z = z$ has density/distribution $g(x, y)$ then

$$g(x) = \int g(x, y)\mathrm{d}y \quad \text{and } g(x|y) = \frac{g(x, y)}{g(y)}$$

density of X given Z respectively X given Y *and* Z.

Apologies for using sloppy $g(x)$, $g(y)$,... notation.

Useful observation II: from conditional distribution to marginal distribution

Suppose $Y|X = x \sim N(c + Ax, V)$ and $X \sim N(\mu, \Sigma)$. Then

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ c + A\mu \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma A^{\mathsf{T}} \\ A\Sigma & A\Sigma A^{\mathsf{T}} + V \end{bmatrix} \right)$$

.

Note: from this and previous useful observation we immediately get

$$U_n | y_{1:(n-1)} \sim N(G\hat{u}_{n-1}, G\Sigma_{n-1}G^{\mathsf{T}} + \Phi)$$

Quick derivation of Kalman filter

Assume we have computed in previous step $\hat{u}_{n-1}$ and Σ_{n-1} (recursion).

Since

$$U_n | y_{1:(n-1)} \sim N(G\hat{u}_{n-1}, G\Sigma_{n-1}G^\mathsf{T} + \Phi)$$

and (by conditional independence)

$$Y_n | u_n, y_{1:n-1} \sim Y_n | u_n \sim N(Fu_n, \Sigma)$$

we get by useful observation II that joint distribution of $(U_n, Y_n)|y_{1:(n-1)}$ is

$$N\left(\begin{bmatrix} G\hat{u}_{n-1} \\ FG\hat{u}_{n-1} \end{bmatrix}, \begin{bmatrix} R_n & R_nF^\mathsf{T} \\ FR_n & FR_nF^\mathsf{T} + \Sigma \end{bmatrix}\right)$$

where $R_n = G\Sigma_{n-1}G^\mathsf{T} + \Phi$.

By useful observation I we obtain $U_n|y_{1:n} \sim U_n|y_{1:(n-1)}, y_n$ as the conditional distribution of U_n given $Y_n = y_n$ derived from the above normal distribution of (U_n, Y_n) given $Y_{1:(n-1)} = y_{1:(n-1)}$.

Hence $U_n|y_{1:n}$ is normal with mean and variance

$$\hat{u}_n = G\,\hat{u}_{n-1} + R_n F^{\mathsf{T}}(F R_n F^{\mathsf{T}} + \Sigma)^{-1}(y_n - FG\,\hat{u}_{n-1})$$
$$\Sigma_n = R_n - R_n F^{\mathsf{T}}(F R_n F^{\mathsf{T}} + \Sigma)^{-1} F R_n$$

Kalman smoother

Suppose we want to compute conditional distribution of U_i given $Y_{1:n}$. This can be done by another recursion backwards in time starting with U_n given $Y_{1:n}$ which we know by now.

Assume that we know (recursion) $U_{i+1}|y_{1:n} \sim N(\tilde{u}_{i+1}, \tilde{\Sigma}_{i+1})$.

We want to compute distribution of $U_i|y_{1:n}$. Condition on U_{i+1} and use conditional independence:

$$U_i|u_{i+1}, y_{1:n} \sim U_i|u_{i+1}, y_{1:i}$$

The conditional distribution $U_i | u_{i+1}, y_{1:i}$ can be derived from the joint distribution $(U_i, U_{i+1}) | y_{1:i}$ which using Kalman filter and useful observation II is

$$N\left(\begin{bmatrix} \hat{u}_i \\ G\hat{u}_i \end{bmatrix}, \begin{bmatrix} \Sigma_i & \Sigma_i G^\mathsf{T} \\ G\Sigma_i & R_i \end{bmatrix}\right).$$

From this we obtain

$$U_i | u_{i+1}, y_{1:n} \sim U_i | u_{i+1}, y_{1:i} \sim N(\hat{u}_i + C_i(u_{i+1} - G\hat{u}_i), H_i)$$

(with $C_i = \Sigma_i G^\mathsf{T} R_i^{-1}$ and $H_i = \Sigma_i - \Sigma_i G^\mathsf{T} R_i^{-1} G\Sigma_i$).

Combining this with $U_{i+1} | y_{1:n} \sim N(\tilde{u}_{i+1}, \tilde{\Sigma}_{i+1})$ we get the desired smoother distribution for U_i:

$$U_i | y_{1:n} \sim N(\hat{u}_i + C_i(\tilde{u}_{i+1} - G\hat{u}_i), C_i \tilde{\Sigma}_{i+1} C_i^\mathsf{T} + H_i)$$

We can now work our way backward in time.

Kalman versus sparse matrix methods

Kalman filter heavily exploits conditional independence of future and past given current state.

Hence restricted to time-series/dynamic models.

Methods based on sparse matrix Cholesky (Miniproject 2) work in any setting with sparse precision matrix for latent Gaussian process.

Exercises

1. show the equivalence of 1.-4. on slide 3.
2. verify the expression for the conditional distribution of X_i on slide 6.
3. check the result regarding the conditional correlation between X_i and X_j below on slide 6.
4. Show that the following three specifications are equivalent:
   4.1 $Y|X = x \sim N(c + Ax, V)$ and $X \sim N(\mu, \Sigma)$
   4.2
   $$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ c + A\mu \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma A^{\mathsf{T}} \\ A\Sigma & A\Sigma A^{\mathsf{T}} + V \end{bmatrix} \right)$$
   4.3 $X \sim N(\mu, \Sigma)$ and $Y = c + AX + \epsilon$ where $\epsilon \sim N(0, V)$ is independent of X.

   (hint: use characteristic function, cf. first lecture)
5. Make an R-implementation of the Kalman filter and smoother for the AR(1)+noise model.