Hierarchical Dynamic Linear Models

André Bonfrer, Michael Braun

January 5, 2012

1 General setup

A k level hierarchical dynamic linear model (HDLM) can be defined as (we define the top of the hierarchy as the kth level, the bottom is the 0th level):

$$\theta_{i-1,t} = F_{it}\Theta_{it} + v_{it} \quad \forall i = 1, \dots, k \tag{1}$$

$$\theta_{k,t} = G_t\Theta_{k,t-1} + H_t + w_t \tag{2}$$

with, typically, $\Theta_{0t} = Y_t$ being some observed outcome (here log sales), $v_{it} \sim N(0, V_{it}, \Sigma)$ and $w_t \sim N(0, W_t, \Sigma)$. Note that we allow for time varying covariance matrixes. The distribution $N(0, V_t, \Sigma)$ is a matrix-variate normal, with left variance V_t and right variance Σ.

Let q represent the number of equations, and r_0 represent the number of parameters at the bottom level of the hierarchy. For ease of exposition, let all response parameters (Θ) and covariance parameters V_{it} and W_t be grouped as Ω and refer to all past data at time t as $D = \{D_1, D_2, \dots, D_T\}$. The likelihood for $Y = \{Y_1, Y_2, \dots, Y_T\}$ (each Y_t is of dimension $r_0 \times q$), for T time periods, can be written as:

$$P(Y \mid D, \Omega) = \prod_{t=1}^{T} 2\pi^{-r_0 q/2} \mid Q_t \mid^{-q/2} \mid \Sigma \mid^{-r_0/2} \exp\left[-\frac{1}{2} tr \left\{(Y_t - f_t)' Q_t^{-1} (Y_t - f_t)\Sigma^{-1}\right\}\right] \tag{3}$$

where (we start by having available, $M_{k,0}$ and $C_{k,0}$ as initial values)

$$f_t = F_{1t}a_{1t} \tag{4}$$

$$a_{it} = F_{i+1,t}a_{i+1} \quad \forall i = 1, \dots, k-1 \tag{5}$$

$$a_{kt} = G_t M_{k,t-1} + H_t \tag{6}$$

$$R_{kt} = G_t C_{k,t-1} G_t' + W_t \tag{7}$$

$$R_{it} = F_{i+1,t}R_{i+1,t}F_{i+1,t}' + V_{i+1,t} \quad \forall i = 1, \dots, k-1 \tag{8}$$

$$Q_t = F_{1t}R_{1t}F_{1t}' + V_{1t} \tag{9}$$

$$E_t = Y_t - f_t \tag{10}$$

$$S_{it} = R_{it}E_{0it}' \quad \forall i = 1, \dots, k \tag{11}$$

$$C_{it} = R_{it} - S_{it}Q_t^{-1}S_{it}' \tag{12}$$

$$
\begin{aligned}
M_{i,t} &= a_{it} + S_{it} Q_t^{-1} E_t \quad \forall i = 1, \ldots, k && (13)\\
S_t &= S_{t-1} + E_t' Q_t^{-1} E_t && (14)\\
n_t &= n_{t-1} + r_0 && (15)
\end{aligned}
$$

The notation E_{ijt} is used to denote a transformation through the levels by premultiplying the lower level F matrix successively by the higher levels. Formally, $E_{ijt} = F_{i+1,t} F_{i+2,t} \ldots F_{j,t},\ i < j$.

The priors for Θ_{k0} and Σ, are jointly distributed as Normal Inverse Wishart $(\Theta_{k0}, \Sigma) \mid D_0 \sim NIW(M_{k0}, C_{k0}, n_0, S_0)$ which in turn is written as (with Σ being of dimension $q \times q$ and X being of dimension $r \times d$:

$$
p(\Theta_{k0}, \Sigma) \propto \mid \Sigma \mid^{-q+(r+d)/2} \times \exp\left\{ -tr[(S + (\Theta_{k0} - M_{k0})' C_{k0}^{-1} (\Theta_{k0} - M_{k0}))\Sigma^{-1}]/2 \right\} \quad (16)
$$

1.1 A case study: application to cross-city wearout effects

The application involves a national brand, determining an allocation of a budget for television advertising expenditure over the entire country. Managers require estimates of how each city's "goodwill" evolves, and measure each city's goodwill. There are two media choices for advertising expenditure, one is national advertising and one is city level (spot TV) advertising.

Our data comprises of J brands, each advertising every week, for T weeks. Advertising expenditure on television is national (there is a more local form of television advertising, known as "spot" advertising, but that usually represents a small fraction of total expenditure for television).

The sales data consists of sales for each brand, each week. Sales are known to be functions of prices, promotions, and seasonal factors (e.g. Van Heerde). The "cross" effects of prices and promotions are often included, to allow for brands to draw customers toward them, from competing brands. In stable categories (e.g. laundry detergents) this translates into market share effects. Advertising also generates cross effects. In addition, we allow for "goodwill" to be generated, representing the baseline attraction of a brand at a point in time.

1.1.1 A non-hierarchical model for decay, wear in and wear out

We begin with the simple version for purposes of illustration as to how the "goodwill" and "advertising quality" metrics evolve over time. We take the case of a single firm using multiple advertising themes (represented by copies, e.g. could be a persuasion theme, a demand stimulation theme, etc.). Following Naik et al (1998), and Bass et al (2008), a brand's advertising (goodwill) stock evolves as a function of the quality and quantity of advertising expenditure (over L themes):

$$
\frac{dB}{dt} = \sum_{l=1}^{L} \tilde{g}(A_{lt}) \kappa_{lt} - \delta B
$$

where κ_{lt} is the the efficacy of an advertising theme at any point in time. This in turn evolves as follows:

$$\frac{d\kappa_l}{dt} = (1 - I(A_{lt}))\delta(1 - \kappa_l) - (c_h + u_l A_{lt})\kappa_l$$

where c_l is a copy wearout and u_l is a "repitition wearout" parameter for theme l, the former being based on the age of the copy, the latter capturing wearout based on extended exposure to the same advertising (therefore a function of ad expenditure). The symbol $I(x)$ is a dummy variable ($= 1$ if $x > 0$ and 0 otherwise). The decay paramater is δ and represents how much "forgetting" there exists. There are two components: one exists if advertising is "on", which is

$$\frac{d\kappa_l}{dt} = -(c_l + u_l A_{lt})\kappa_l$$

the other is when advertising is "off":

$$\frac{d\kappa_l}{dt} = \delta(1 - \kappa_l) - c_l\kappa_l$$

Adding in a component for the random shock to the evolution equation above, this can be set up as a dlm (non-hierarchical) with:

$$\begin{aligned} Y_t &= F_t\Theta_{1t} + \mathbf{X}_t\beta + v_{1t} \\ \Theta_{1t} &= G_t\Theta_{1t-1} + H_t + \mathbf{w}_t \end{aligned} \tag{17}$$

where β is a $P \times 1$ vector of parameters (non time varying)[1], $\mathbf{X}_t$ is a $1 \times P$ set of covariates (e.g. effect of price or promotions). The random terms are $v_{1t} \sim N(0, V_t)$, $w_t \sim N(0, W_t)$ and with the evolution matrix (G_t):

$$G_t = \begin{bmatrix} (1-\delta) & \tilde{g}(A_{1t}) & \dots & \tilde{g}(A_{Lt}) \\ 0 & (1 - c_1 - u_1 A_{1t}) - \delta(1 - I(A_{1t})) & \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \dots & (1 - c_L - u_L A_{Lt}) - \delta(1 - I(A_{Lt})) \end{bmatrix} \tag{18}$$

and

$$H_t = \begin{bmatrix} 0 \\ \delta(1 - I(A_{1t}) \\ \vdots \\ \delta(1 - I(A_{lt}) \\ \vdots \\ \delta(1 - I(A_{Lt}) \end{bmatrix} \tag{19}$$

[1]The case of multiple firms, or multiple brands, would require Y_t to be a vector. Consequently, F_t would need to be a matrix with a corresponding block-diagonal structure. This is also true if multiple, simultaneous equations are being modeled (e.g. advertising)

Finally, F_t is of dimension $1 \times (L+1)$ which is $F_t = [1 \;\; 0 \; \dots \; 0]$. This captures a latent structure for the "intercept", or goodwill as a function of advertising.

With F_t substituted in (17), allows sales to be a function of goodwill, plus covariates. Goodwill (B_t) evolves as a function of advertising quality of each of the themes, multiplied by the expenditure on each theme:

$$B_t = (1-\delta)B_{t-1} + \sum_{l=1}^{L} \tilde{g}(A_{lt})\kappa_{lt-1} + w_{0t}$$

and the quality of each advertising theme evolves as:

$$\kappa_{lt} = [(1 - c_l - u_l A_{lt}) - \delta(1 - I(A_{lt}))]\,\kappa_{lt-1} + \delta(1 - I(A_{lt})) + w_{lt} \quad (20)$$

The additive random terms are stacked into $\mathbf{w}_t = \{w_{0t}, w_{1t}, \dots, w_{Lt}\} \sim N(0, W_t)$.

1.1.2 Hierarchical model

The hierarchical model is different from the non-hierarchical model, in that we have a number of cities for which we have the same data as above. We assume that most of the parameters are city-specific, and are drawn from a distribution across cities. We now have J competing brands that all advertise, though not always simultaneously. In the below model, we simplify to allow only one theme per brand (here we refer to this as total national advertising).

Each of the brands has a "quality" of its advertising, κ_{jk}, here defined as how much a dollar spent by brand k can affect the goodwill of brand j. When $j \neq k$, this is analagous to the cross-competitive effect often used in sales response models, but here applies to the quality of that advertising expenditure.

So now the evolution equation for brand j is a function of advertising of all competing brands $k \in \{1, \dots, J\}$ via interactions between qualities and corresponding expenditures:

$$B_{jt} = (1-\delta)B_{jt-1} + \sum_{k=1}^{J} \tilde{g}(A_{kt})\kappa_{jkt-1} + w_{0jt} \quad (21)$$

The quality component of each brand's advertising evolves as:

$$\kappa_{jkt} = ((1 - c_k - u_k A_{kt}) - \delta(1 - I(A_{kt})))\,\kappa_{jkt-1} + \delta(1 - I(A_{kt})) + w_{jkt} \quad (22)$$

The way to read this is, κ_{ijt} is the quality of j's advertising relevant to the goodwill of brand i. This allows us to capture the direct effect of competition, on the sales of the focal brand. The competitive effect in this specification is on goodwill, so we expect this to be negative. The stochastic component to do with the evolution of the parameters is now assumed to be drawn from a matrix-normal distribution:

$$\mathbf{w}_t = \begin{bmatrix} w_{01t} & \dots & w_{0Jt} \\ w_{11t} & \dots & w_{1Jt} \\ \vdots & & \\ w_{j1t} & \dots & w_{jJt} \\ & \dots & \end{bmatrix} \sim N(0, W_t, \Sigma) \quad (23)$$

The notation $N(M, C, \Sigma)$ introduces a matrix-normal distribution, which has a left variance C and right variance of Σ, and is equivalent to the multivariate normal $N(vec(M), \Sigma \otimes C)$.

This gives us the structure below:

$$\begin{aligned} Y_t &= F_{1t}\Theta_{1t} + \mathbf{X}_t\beta + v_{1t} \\ \Theta_{1t} &= F_{2t}\Theta_{2t} + v_{2t} \\ \Theta_{2t} &= G_t\Theta_{2t-1} + H_t + w_t \end{aligned} \tag{24}$$

The observed outcome data is contained in Y_t which is a $I \times J$ matrix containing sales for each brand (contained in the jth column of Y_t and each city (ith row). Accordingly, F_{1t} is a $I \times I(J+1)$ matrix, Θ_{1t} is dimension $I(J+1) \times J$.

The parameters contained in Θ_{2t}, can be thought of as mean level parameters for the top level of the hierarchy. Θ_{2t} is now a J column matrix (dimension is $(J+1) \times J$):

$$\Theta_{2t} = \begin{bmatrix} B_{1t} & \dots & B_{Jt} \\ \kappa_{11t} & \dots & \kappa_{J1t} \\ & \vdots & \\ \kappa_{1Jt} & \dots & \kappa_{JJt} \end{bmatrix} \tag{25}$$

The next level in the hierarchy maps the mean levels to the city specific levels. We have I cities of data. In this regard, F_{2t} is able to make each parameter in the top level equation a function of either covariates or just of the mean level across cities. For example, specifying F_{2t} as a stacked matrix of I identity matrixes, each of dimension $(J+1) \times (J+1)^2$:

$$F_{2t} = \{\mathbf{1}_{J+1}; \dots; \mathbf{1}_{J+1}\}$$

which is of dimension $I(J+1) \times (J+1)$. This makes each row of the Θ_{2t} parameter a mean value, with a deviation for each city. We now have goodwill parameters (B_{ijt}) estimated for each brand and each city i. Eg. goodwill for brand j for city i is:

$$B_{ijt} = B_{jt} + v_{2t}$$

with $v_{2t} \sim N(0, V_2, \Sigma)$.

The evolution matrix now only specified on the mean levels (across cities) for the components of advertising. This is the same as (18) but now the evolution occurs across brands' advertising:

$$G_t = \begin{bmatrix} (1-\delta) & \tilde{g}(A_{1t}) & \dots & \tilde{g}(A_{Jt}) \\ 0 & (1 - c_1 - u_1 A_{1t}) - \delta(1 - I(A_{1t})) & \dots & 0 \\ \vdots & 0 & \ddots & \\ 0 & 0 & & (1 - c_J - u_J A_{Jt}) - \delta(1 - I(A_{Jt})) \end{bmatrix} \tag{26}$$

[2]Making each parameter a function of covariates will obviously change this structure.

This framework could be used in a number of ways. For example, we can include different themes in the response functions, and study how best to allocate expenditures given some allocation of themes. Alternatively, the same theme could be studied from a competitive effects perspective.

The specification above will be included in (16), and we combine multivariate normal specifications for these parameters, across brands. Estimation of the wearin and wearout parameters is done using adaptive rejection sampling. The remainder of the parameters can be estimated using Gibbs steps.