Implementation notes for HDLM ad spend model

André Bonfrer and Michael Braun

February 3, 2015

Definitions of symbols

1 Two-level hierarchical model

Assume we have outcome (e.g sales) data for J brands, in N cities, and observe these data over T time periods. We represent this outcome (Y_t) at time t as a $N \times J$ matrix, with the rows representing city level observations, and columns representing brand sales (outcome) data. We have a hierachy of city level at the lowest (each city has its own sales) and at the highest level we have dynamics at the mean level (e.g. mean prices, or the effects of national level advertising). This is written as:

$$\begin{aligned} Y_t &= F_{11t}\Theta_{11t} + F_{12t}\Theta_{12} + v_{1t} && (1)\\ \Theta_{11t} &= F_{2t}\Theta_{2t} + v_{2t} && (2)\\ \Theta_{2t} &= \tilde{G}_t\Theta_{2,t-1} + \tilde{H}_t + w_t && (3) \end{aligned}$$

The components that affect each city's sales directly, are in the F_{12t} matrix, with a corresponding non-time varying coefficient matrix. The time varying component at the city level is contained in the $F_{11t}\Theta_{11t}$ component. In addition we have an innovation function (sometimes called a control variable) in the evolution equation, H_t. This component shifts elements of Θ_{2t} but is not relative to it.

We use a matrix normal distribution for all covariance terms:

$$\begin{aligned} v_{lt} \mid \Sigma, V_l &\sim N(0, V_l, \Sigma) && (4)\\ w_t \mid \Sigma, W &\sim N(0, W, \Sigma) && (5) \end{aligned}$$

Each matrix normal distribution has a left and right variance matrix, e.g. V_1, Σ respectively. The right variance governs (column) cross equation covariatance. The left variance captures row covariance, which is either concurrent (V_l) or time based variation (W). The left variance V_1 represents variation across cities. The left variance V_2 represents concurrent variance across mean values for different state variables. We can simplify notation a bit by using a set $\Psi = \{V_1, V_2, W\}$.

At the first level we have $\bar{Y}_t = Y_t - F_{12t}\Theta_{12}$, which does not have a hierarchical counterpart (i.e. homogenous response to covariates contained in F_{12t}). The Θ_{11t} component then has both time varying and non-time varying heterogeneous responses. We can rewrite our HDLM as:

$$\begin{aligned} \bar{Y}_t &= F_{11t}\Theta_{11t} + v_{1t} && (6)\\ \Theta_{11t} &= F_{2t}\Theta_{2t} + v_{2t} && (7)\\ \Theta_{2t} &= \tilde{G}_t\Theta_{2,t-1} + \tilde{H}_t + w_t && (8) \end{aligned}$$

We use the tilde ($\tilde{\cdot}$) in the above to represent intermediate variables (those that depend on other parameters).

Data likelihood

We write the data likelihood, recognising that simple substitutions can be made to take care of the time invariant and homogenous parameters. Let D_{t-1} represent all information (including state variables) available at time t. So D_0 is initial information about the states and priors. At any time period, $\Sigma \mid D_{t-1}$ is distributed as an Inverse Wishart ($IW(\nu_{t-1}, \Omega_{t-1})$. For any time period, the joint density of the data Y_t and Σ is a matrix normal inverse Wishart (or a product of a matrix normal with inverse Wishart):

$$P(\bar{Y}, \Sigma \mid \mathcal{D}_0, \Psi) = \prod_{t=1}^{T} P(\bar{Y}_t \mid \Sigma, D_{t-1}, \Psi) P(\Sigma \mid D_{t-1}, \Psi) \tag{9}$$

$$= \prod_{t=1}^{T} (2\pi)^{-\frac{NJ}{2}} |Q_t|^{-\frac{J}{2}} |\Sigma|^{-\frac{N}{2}} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\left(\bar{Y}_t - f_t\right)' Q_t^{-1} \left(\bar{Y}_t - f_t\right) \Sigma^{-1}\right)\right] \times IW(\nu_{t-1}, \Omega_{t-1}) \tag{10}$$

Integrating out Σ (see our technical appendix on matrix T) gives the following data likelihood:

$$P(\bar{Y}|\cdot) = \prod_{t=1}^{T} P(\bar{Y}_t | y_{1:t-1}, \cdot)$$

$$= \mathcal{K} \left(\prod_{t=1}^{T} |Q_t|^{-\frac{J}{2}}\right) |\Omega_0 + \sum_{t=1}^{T} \left(\bar{Y}_t - f_t\right)' Q_t^{-1} \left(\bar{Y}_t - f_t\right)|^{-\frac{\nu_0 + TN}{2}} \tag{11}$$

eq:LL-T

where:

$$\mathcal{K} = \pi^{-\frac{NJT}{2}} \frac{\Gamma_J\left(\frac{\nu_0 + TN}{2}\right)}{\Gamma_J\left(\frac{\nu_0}{2}\right)} |\Omega_0|^{-\frac{\nu_0}{2}}$$

2 Theoretical development of the dynamic hierarchical effects of advertising

In our application (FMCG), we take the perspective of a national-level marketing manager observing city level observations of sales and prices, allocating an advertising expenditure over time and across cities. We focus on the effectiveness of network advertising, which as a unit of analysis is observed only at a national level. That is, each city is exposed to the same network television patterns. Effectiveness is a dynamic function of the content in the creatives used in the campaign. At any time, there are a number of distinct creatives observed in the market place. The manager can choose to spend on existing/past creatives, or to invest in and launch a new creative. The industry setting is such that there are a total of J competitors. Expenditure allocation decisions are made on a continuous time basis but we use weekly level sales data so we convert the expenditures to a weekly level.

In the full competitive model, all covariates (advertising, price and promotions) have both an own and cross effect. In the matrix normal set up of the HDLM above, this is automatically specified by having a matrix normal of the brand sales in the columns, and the covariates are each brands' covariates.

In addition to competitive effects and competitive interference effects, the effectiveness of the advertising campaign wears out over time. This is because of the wearout of individual messages. In line with past work (e.g. Naik et al 1998, Bass et al 2008, Braun and Moe 2012), we think of ad effectiveness for each brand j creative m, denoted by q_{mjt}.

We abstract from any individual message and start with the standard Nerlove-Arrow type evolution of "brand" or overall "advertising" effectiveness (defined as the ability for a dollar spent on advertising to lift sales volume), represented by:

$$B_{jt} = (1 - \delta) B_{jt-1} + q_{jt-1} g(A_{ijt}) \tag{12}$$

eqn:bqj1

where $g(\cdot)$ is some transformation of advertising ($= 0$ if $A_{ijt} = 0$). We will return to the issue of including competitive effects of a focal brand's advertising, multiple media messages and the effect of competing

brands' advertising on the focal brand's sales.

One way to build dynamic advertising effectiveness is by using a differential equation with respect to time, of brand j's advertising:

$$\frac{dq_j}{dt} = -a(A_j)q_j + (1 - \mathbb{I}(A_j))\delta(1 - q_j)$$

If advertising is positive, then it wears out as a function of $a()$ given above (e.g. $a(A) = c + wA$). If advertising is turned off, then there is a rejuvenation effect given by $\delta(1 - q_j)$. Naik et al (1998) point out that, if advertising is turned off, q_{jt} tends to the following value:

$$\lim_{t\to\infty} q_{jt} = \frac{\delta}{c + \delta}$$

The evolution of advertising effectiveness is then (let $dq = q_t - q_{t-1}$ and $dt = 1$):

$$\begin{aligned} q_{jt} - q_{jt-1} &= -a(A_j)q_{jt-1} + (1 - \mathbb{I}(A_j))\delta(1 - q_{jt-1}) \\ q_{jt} &= q_{jt-1} - a(A_j)q_{jt-1} + (1 - \mathbb{I}(A_j))\delta(1 - q_{jt-1}) \\ &= q_{jt-1}\left[1 - a(A_j) - \delta(1 - \mathbb{I}(A_j))\right] + \delta(1 - \mathbb{I}(A_j)) \end{aligned} \tag{13}$$

A few notes about this. First, it is only valid if $q > 0$. For the cross effectiveness of advertising in our competitive model, we need to allow for this to be negative. If $q < 0$ then we need a different dynamic:

$$\frac{dq_j}{dt} = -a(A_j) - (1 - \mathbb{I}(A_j))\delta(1 + q_j)$$

which allows advertising effectiveness to return to a larger negative value if advertising for that brand is switched off. If advertising is switched off, it has the same asymptote as before, but negative in value[1]:

$$\lim_{t\to\infty} q_{jt} = -\frac{\delta}{c + \delta}$$

This leads to the following:

$$\begin{aligned} q_{jt} - q_{jt-1} &= -a(A_j)q_{jt-1} - (1 - \mathbb{I}(A_j))\delta(1 + q_{jt-1}) \\ q_{jt} &= q_{jt-1} - a(A_j)q_{jt-1} - (1 - \mathbb{I}(A_j))\delta(1 + q_{jt-1}) \\ &= q_{jt-1}\left[1 - a(A_j) - \delta(1 - \mathbb{I}(A_j))\right] - \delta(1 - \mathbb{I}(A_j)) \end{aligned} \tag{14}$$

The only difference between (13) and (14) is the "innovation" component ($sign(q)\delta(1 - \mathbb{I}(A_j))$) which depends on the sign of the advertising effectiveness component. See below for how we include the innovation component (it does not enter $\tilde{G}_t$).

Second, we want to incorporate new creatives. For this, we propose the differential equation:

$$\frac{dq_j}{dt} = -a(A_j)q_j + (1 - \mathbb{I}(A_j))\delta(1 - sign(q_j)q_j) + sign(q_j)\phi_j E_j$$

where E_j is some metric for new creatives. E.g. it could count the number of new creatives being used (or the proportion of the creatives that have not been shown before).

2.1 Specifying Θ_{2t}

The matrix Θ_{2t} is a (dense) matrix of time varying parameters (also called state variables). Without the $P \times J$ matrix of time varying parameters, the rows of this correspond to

[1] Our issues with the identification of c means we set this to 0. The obvious implication is that the asymptotic value of q is then equal to one.

$$\Theta_{2t} = \begin{bmatrix} B_{1t} & B_{2t} & \dots & B_{Jt} \\ q_{11t} & q_{21t} & \dots & q_{J1t} \\ q_{12t} & q_{22t} & \dots & q_{J2t} \\ \vdots & & & \\ q_{1Jt} & q_{21t} & \dots & q_{JJt} \end{bmatrix} \tag{15}$$ eqn:t2a

with F_{2t} being the matrix (dimension $N \times (J+1)$) that translates these states to the city level. For example, corresponding to the above:

$$F_{2t} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \vdots & & & \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{16}$$ eqn:t2b

The zero elements of this matrix make the ad effectiveness parameters/states latent, with their role only being played in the evolution matrix below. For (15) and (16) they combine to provide a $N \times J$ matrix which just depends on the intercept (B_{jt} in 12). eqn:t2a eqn:t2b eqn:bqj1

2.1.1 Adding time varying effects of other covariates

To add time varying effects of other covariates, we add in rows corresponding to each effect to Θ_{2t}, e.g. adding J rows[2], a price "mean" parameter for each price variable (i.e. this is the mean value for the price elasticity across cities):

$$\Theta_{2t} = \begin{bmatrix} B_{1t} & B_{2t} & \dots & B_{Jt} \\ q_{11t} & q_{21t} & \dots & q_{J1t} \\ q_{12t} & q_{22t} & \dots & q_{J2t} \\ \vdots & & & \\ q_{1Jt} & q_{21t} & \dots & q_{JJt} \\ \theta^p_{11t} & \theta^p_{21t} & \dots & \theta^p_{J1t} \\ \theta^p_{12t} & \theta^p_{22t} & \dots & \theta^p_{J2t} \\ \vdots & & & \\ \theta^p_{1Jt} & \theta^p_{21t} & \dots & \theta^p_{JJt} \end{bmatrix} \tag{17}$$

[2]This is a bit confusing, but because we have cross effects for each covariate we add, then adding in just price adds $P = J$ covariates. If we allow two types of covariates (e.g. price and promotion) then $P = J + J$

Then F_{2t} becomes (with dimension $N(1+P) \times (1+J+P)$) :

$$F_{2t} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \vdots & & & & & & & \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \vdots & & & & & & & \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{18}$$

2.2 Specifying F_{11t}

The city level matrix of F_{11t} is the mean for the distribution of $\bar{Y}_t$ as a (simple additive) function of the covariates (and latent space) at the city level. Without any additional time varying effects of covariates, this is an $N \times N$ identity matrix:

$$F_{11t} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{19}$$

The addition of city specific price covariates, for which the effects vary over time, then the matrix is $N \times N(1+P)$ where P corresponds to the number of covariates (times J). For example, consider adding price for J brands, we will add $P = J$ covariates to F_{11t}:

$$F_{11t} = \begin{bmatrix} 1 & p_{11t} & p_{1jt} & \dots & p_{1Jt} \dots & 0 & & & & \dots & 0 \\ \vdots & & & & & & & & & & \\ 0 & 0 & 0 & 0 & \dots & 1 & p_{n1t} & p_{njt} & \dots & p_{nJt} & \dots & 0 \\ \vdots & 0 & \ddots & & & & & & & & & \end{bmatrix} \tag{20}$$

where p_{njt} is price in city n for brand j at time t. In the above, the dimension would be $N \times N(1+P)$.

2.3 Specifying $\tilde{G}_t$

The corresponding "evolution" matrix $\tilde{G}_t$ is $(J+1+P) \times (J+1+P)$, and is upper triangular. We illustrate this below by ignoring the $\tilde{G}_t$ component for the P time varying components (which would just be an identity matrix of dimension $P \times P$). Let $\tilde{g}(A_{jt})$ be some transformation function of ad spend for brand j at time t. The function $\mathbb{I}(\cdot)$ is an indicator function that takes a value of 1 if the argument is greater than zero, and 0 otherwise.[3]

[3]If we allow for different "cross" effect wearouts, then we have an evolution matrix for each brand. We would no longer be able to use the matrix normal T distribution for this, and would have the following jth evolution matrix premultiplying the jth column of

$$Here\ \tilde{G}_t = \begin{bmatrix} (1-\delta) & \tilde{g}(A_{1t}) & \dots & \tilde{g}(A_{Jt}) \\ 0 & (1-c_1-u_1A_{1t})-\delta(1-\mathbb{I}(A_{1t})) & \dots & 0 \\ \vdots & 0 & \ddots & \\ 0 & 0 & & (1-c_J-u_JA_{Jt})-\delta(1-\mathbb{I}(A_{Jt})) \end{bmatrix} \tag{22}$$

Computing H_t

The innovation component adds an amount to each state parameter, and for each brand. Therefore the dimension of the matrix $\tilde{H}_t$ is $(1+J+P)\times J$, and depends on ad spend and the parameter δ. Again ignoring any additional time varying effects of covariates (so $P=0$) we have:

$$H_t^1 = \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ sign(q_{11})\delta(1-\mathbb{I}(A_{1t})) & \dots & sign(q_{j1})\delta(1-\mathbb{I}(A_{1t})) & \dots & sign(q_{J1})\delta(1-\mathbb{I}(A_{1t})) \\ \vdots & \ddots & & & \\ sign(q_{1j})\delta(1-\mathbb{I}(A_{jt})) & sign(q_{kj})\delta(1-\mathbb{I}(A_{jt})) & & & \\ \vdots & & & \ddots & \\ sign(q_{1J})\delta(1-\mathbb{I}(A_{Jt})) & \dots & & & sign(q_{JJ})\delta(1-\mathbb{I}(A_{Jt})) \end{bmatrix} \tag{23}$$

$$H_t^2 = \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ signq_{11}\phi_{11}E_{1t} & signq_{21}\phi_{21}E_{1t} & \dots & & signq_{J1}\phi_{J1}E_{1t} \\ \vdots & \ddots & & & \\ signq_{1j}\phi_{1j}E_{jt} & \dots & signq_{jj}\phi_{jj}E_{jt} & \dots & signq_{Jj}\phi_{Jj}E_{jt} \\ \vdots & & & & \\ signq_{1J}\phi_{1J}E_{Jt} & signq_{2J}\phi_{2J}E_{Jt} & \dots & & signq_{JJ}\phi_{JJ}E_{Jt} \end{bmatrix} \tag{24}$$

Then $\tilde{H}_t = H_t^1 + H_t^2$. Once again, there will be an additional P rows to $\tilde{H}_t$ corresponding to additional time varying covariates.

Iterative estimation

To estimate the likelihood in Equation (11), we need to compute $\bar{Y}_t$, f_t, and Q_t. The matrix Y_t is the observed, dependent variable, so we need to get f_t and Q_t, and $\bar{Y}_t = Y_t - F_{12t}\Theta_{12}$. Conditional on estimates from time $t-1$, and all data and prior information, we can follow the following algorithm at time t.

1. Compute $\tilde{G}_t$ using A_t, c, u, ϕ and δ, then $\tilde{G}_t$ to include any additional time varying effects for covariates. Similarly, create $\tilde{H}_t$ including $P\times J$ matrix of zeros.
2. Set $a_{2t} = \tilde{G}_t M_{2,t-1} + \tilde{H}_t$.
3. Set $f_t = F_{11t}F_{2t}a_{2t}$.
4. Set $R_{2t} = \tilde{G}_t C_{2,t-1}\tilde{G}_t' + W$

Θ_{2t}:

$$\tilde{G}_{jt} = \begin{bmatrix} (1-\delta) & \tilde{g}(A_{1t}) & \dots & \tilde{g}(A_{Jt}) \\ 0 & (1-c_{1j}-u_{1j}A_{1t})-\delta(1-\mathbb{I}(A_{1t})) & \dots & 0 \\ \vdots & 0 & \ddots & \\ 0 & 0 & & (1-c_{Jj}-u_{Jj}A_{Jt}-\delta(1-\mathbb{I}(A_{Jt})) \end{bmatrix} \tag{21}$$

5. Set $R_{1t} = F_{2t} R_{2t} F_{2t}' + V_2$
6. Set $Q_t = F_{11t} R_{1t} F_{11t}' + V_1$
7. Set $S_{2t} = R_{2t} \left[F_{11t} F_{2t}\right]'$
8. Set $M_{2t} = a_{2t} + S_{2t} Q_t^{-1} (\bar{Y}_t - f_t)$
9. Set $C_{2t} = R_{2t} - S_{2t} Q_t^{-1} S_{2t}'$

Then, iterate over t to estimate the data likelihood. Note that the homogenous time invariant component at level 1 of the hierarchy is handled by the transformed variable, $\bar{Y}_t$ which appears in the posterior and can be numerically estimated.

Data structures

The matrices A ($T \times J$) and F_{12t} are standard, dense covariate structures. The matrix F_{11t} is given above, and is sparse. Similary with F_{2t}. We separated out the time-invariant homogenous effects in $F_{12t}\Theta_{12}$, so the matrix F_{12t} is $N \times K$. These are non hierarchical and non time varying in their effects.

Specifying parameters

Table 1 summarizes the parameters that need to be estimated, assuming that the covariance matrices are stationary. The number of parameters lists is the number of *unique* elements. For example, in a symmetric matrix there are, at most, $k(k+1)/2$ unique parameters. But this can still be a large number, so we should think about some kind of dimensionality reduction.

To estimate these parameters, we should transform them all to be unbounded. Otherwise, we need to modify the GDS algorithm to handle constrained optimization and simulation (which is possible, but tedious and uninteresting).

For the dense cases of V_1, V_2 and W, typically we would estimate the elements of the lower Cholesky decomposition (taking logs of the diagonal elements to ensure that they are positive). If we add structure to those matrices, we need to reconsider the transformation. However, block diagonals should still allow us to use the Cholesky decomposition approach.

Priors

We need to choose initial values for M_{20} and C_{20}. What do these matrices mean, intuitively? They indicate prior information about the starting states (the means across cities). This could come from theory, another process, or could be made quite diffuse. For example, the way we specify the ad effectiveness (above) could provide us with some prior on the initial state that is constrained to be close to 0. Similary, our understanding of price sensitivity is around -1 to -2 so we could provide such a information as a prior. Of particular interest is the "new" creative effectiveness. The overall effect of this should be somewhat proportional to the effect of the base campaign.

Prior on δ: if δ really is between 0 and 1, we could make the prior uniform. But is that realistic? We probably want a density that places zero probability at 0 and 1, or at least 0. Does the literature give us any prior information about what this decay parameter should be? Yes, it says that δ is usually around $0.1 - 0.2$ (for weekly data). We could use a beta$(1,3)$ as a prior. One problem this may raise (in the GDS) is that the parameter is constrained between zero and one.

Priors on $c_{1:J}$ and $u_{1:J}$: depends on the domain. Are they all between 0 and 1. Also, could they be correlated? Would it make sense that if one brand had high wearout, another brand could as well? Do we have prior information on this? Again, c depends on the ad effectiveness value, but is unlikely to be much different from 0. The w value depends on the scale used for advertising expenditure and is expected to be

positive. The ϕ parameters are likely to be small since it is unlikely any one creative can have a substantial impact on the overall effectiveness of the campaign.

Priors on V_1, V_2 and W: first, we need better intuition about what these matrices represent. Then, we can come up with a range of reasonable values for the parameters. Given the complexity of the model, we will need to regularize it with prior information. And it would be good to give these priors careful though. Way too many marketers are careless with their "uninformative" priors.

Prediction

Once we estimate these top-level parameters, we might want to simulate data. That means we need posterior predictive distributions of Y. Can we do that without simulating the Θ parameters directly? Note that we do not collect any Θ draws during the estimation process, since they are all integrated out. We should be able to use forward filtering, backward sampling (possibly with smoothing) to obtain the city level state variables.

2.4 Focus on initial qualities

The accommodation of initial qualities could be a possible extension. Several possible outcomes could have some interesting implications. First, consider if the distribution across creatives is quite broad, i.e. there is wide heterogeneity in the quality of ads drawn (might want to discuss some of the creative agency literature). This would likely have a quite different outcome in terms of the decision to draw new ads (and the number of creatives being displayed) than if the heterogeneity is quite low. As a "hunch" we may expect the manager to want to try more creatives, sticking with the ones that seem to be high quality, but drawing new creatives if it is of low quality. We may also expect that the pattern of expenditure is quite different - with more pulsing behavior if the ad is high quality, but perhaps some initial pattern to 'learn' about the quality of the advertisement once introduced. Need to also consider how pre-testing of creatives may change these results.

The problem with the estimation of this is in the size of the evolution matrix required which for just the advertising part (instead of a $J+1$ by $J+1$ matrix) requires a matrix which is of dimension equal to (with rather sloppy notation) $m+1$ by $m+1$ where m is the total number of creatives across brands. We start by seeing if we can collapse this across creatives somehow.

2.4.1 Additional to be done:

The most important is to include interference effects for each advertising component.

- Analytics on how to launch ads over time (constant noise). Each ad has fixed cost and variable cost of creating. The variable cost is assumed measured by $g(\cdot)$.
- How much does noise impact the pattern of spending for a focal brand?
- Analytically derive the optimal number of creatives to run at any time, and look at how this depends on the variability in inital quality. Perhaps consider a choice of creative media agency? One with high variability, but low average quality, one with low variability? If it's a beta distribution we could also look at the media agencies that sometimes produce a hit, sometimes a complete flop, versus one that is more reliable "average" quality.
- Advanced - investigate optimal "pulsing" strategy

Symbol	Definition	Dimension
$l = 1 \dots L$	hierarchy level index (level 1 is city, level 2 is national)	
$j = 1 \dots J$	brand index	
$i = 1 \dots N$	city index	
$t = 1 \dots T$	time index	
$k = 1 \dots K$	time invariant covariate index at level 1	
$p = 1 \dots P$	time varying covariate index at level 1	
	Observed data	
Y_t	dependent variable. Sales by city and brand	$N \times J$
F_{11t}	Data ('explanatory variables') matrix. Sparse 0/1	$N \times N(1+P)$
F_{12t}	Data ('explanatory variables') matrix. Sparse 0/1	$N \times K$
F_{2t}	Hierarchical matrix.	$N(1+P) \times (1+J+P)$
A_t	ad spend by city, brand	$N \times J$
F_{12t}	covariate matrix for the 1st hierarchy level (by city, time varying, but observed)	$N \times P$
	Prior parameters (predetermined)	
M_{20}	some prior matrix-normal parameter for Θ_{20}	$(1+J+P) \times J$
C_{20}	some prior matrix-normal parameter for Θ_{20}	$(1+J+P) \times (1+J+P)$
Ω_0	prior inverse Wishart scale parameter matrix for Σ	$J \times J$
ν_0	prior inverse Wishart degrees of freedom parameter for Σ	scalar
	other hyperprior parameters go here	
	parameters that need to be estimated	
Θ_{12}	time invariant coefficients on city and brand specific covariates, non time varying (covariates are F_{12t})	$K \times J$
Θ_{11t}	time varying coefficients on city and brand specific covariates (covariates are F_{11t})	$N(1+P) \times J$
Θ_{2t}	time varying (mean) coefficients (covariates can be introduced in F_{2t})	$(1+J+P) \times J$
c	vector of copy wearout parameters, brand-specific	J
u	vector of ad wearout parameters, brand-specific	J
ϕ	matrix of ad wearout parameters, brand-specific	$J \times J$
δ	decay parameter, $0 < \delta < 1$	scalar
V_1	city level covariance matrix	$N \times N$
V_2	hierarchical covariance matrix (across model components)	$N(1+P) \times N(1+P)$
W	evolution covariance matrix (over time)	$(1+J+P) \times (1+J+P)$
	Intermediate terms	
$\tilde{G}_t$	evolution matrix (upper triangular)	$(1+J+P) \times (1+J+P)$
$\tilde{H}_t$	matrix depending on δ and A_t	$(1+J+P) \times J$
a_{2t}	mean for prior distribution of $\Theta_{2t} \mid D_{t-1}$	$(1+J+P) \times J$
f_t	mean of predictive distribution for $Y_t \mid D_{t-1}$	$N \times J$
Q_t	covariance of predictive distribution for $Y_t \mid D_{t-1}$	$N \times N$
R_{1t}	covariance for prior distribution of $\Theta_{1t} \mid D_{t-1}$	$N(1+P) \times N(1+P)$
R_{2t}	covariance for prior distribution of $\Theta_{2t} \mid D_{t-1}$	$(1+J+P) \times (1+J+P)$
M_{2t}	mean of posterior distribution $\Theta_{2t} \mid D_t$	$(1+J+P) \times J$
C_{2t}	covariance for posterior distribution $\Theta_{2t} \mid D_t$	$(1+J+P) \times (1+J+P)$
S_{2t}	used to translate variance from previous level	$(1+J+P) \times N$

Symbol	Note	Num Pars (if dense)	Reduce by...	reduced parameters
V_1	symmetric pos-def	$N(N+1)/2$	make diagonal	N
			spatial structure	something $> N$
V_2	symmetric pos-def	$[N^2(1+P)^2 + N(1+P)]/2$	block diagonal?	$N(2+P(3+P))/2$
W	symmetric pos-def	$(1+J+P)(2+J+P)/2$	diagonal	(1+J+P)
			block diagonal	1 + J(J+1)/2 + P(P+1)/2
δ	scalar between 0 and 1	1		
ϕ	dense matrix	$J \times J$	make symmetric	J(J+1)/2
c, u	vector, likely positive and close to zero, and may be negative	J each		
Θ_{12}	time invariant homogenous coefficient matrix	$K \times J$	no intercept	

tab:parameters

Table 1: Parameters to be estimated