# Quantitative Trendspotting

Rex Yuxing Du and Wagner A. Kamakura

## Web Appendix A – Inferring and Projecting the Latent Dynamic Factors

The procedure for inferring the latent state variables (i.e., $z_t \stackrel{\text{def}}{=} [\alpha_t \quad \beta_t \quad \delta_t \quad \gamma_t \quad \dots \quad \gamma_{t-s-2}]'$), from the observed indicators (i.e., $y_t$ from t = 1 through T), given the model parameters (i.e., $B$, $L$, $\Sigma_u$, $\Sigma_\varepsilon$, $\Sigma_\eta$, $\Sigma_\zeta$, and $\Sigma_\xi$ in Equations 1, 2, and 2.1 through 2.4), consists of two passes through the data: a forward pass (periods 1 through T) referred to in the literature as Kalman filter, and a backward pass (periods T through 1) often referred to as Kalman smoother (cf. Shumway and Stoffer 2000). Below we provide details on this inference procedure, which is important for understanding the mapping from the observed indicators to the latent state variables, which in turn determine the latent dynamic factors ($f_t$).

<u>*Kalman filter (forward pass)*</u>[1]

Step 1 – initialize $z_{0|0}$ and $var(z_{0|0})$, i.e., prior of the initial state at period 0

- $z_{0|0} = \mu_0$
- $var(z_{0|0}) = \Sigma_0$

Repeat Steps 2 and 3 for t = 1, …, T

Step 2 – calculate $z_{t|t-1}$ and $var(z_{t|t-1})$ i.e., expectation and uncertainty about the state in t given data observed up until t-1

- $z_{t|t-1} = Cz_{t-1|t-1}$

[1] The parametric definitions of matrices $\mu_0$, $\Sigma_0$, A, B, C, $\Sigma_u$ and $\Sigma_v$ that appear in the procedure described below are given in Equations B1 and B2 in Web Appendix B.

- $var(z_{t|t-1}) = Cvar(z_{t-1|t-1})C' + \Sigma_v$

Step 3 – calculate $z_{t|t}$ and $var(z_{t|t})$ i.e., expectation and uncertainty about the state in t given data observed up until t

- $K_t = var(z_{t|t-1})A'[Avar(z_{t|t-1})A' + \Sigma_u]^{-1}$ (a.k.a. Kalman gain from t)
- $z_{t|t} = z_{t|t-1} + K_t[y_t - Az_{t|t-1} - B]$ (additional data from t used to update expectation)
- $var(z_{t|t}) = [I - K_tA]var(z_{t|t-1})$ (additional data from t used to reduce uncertainty)

*Kalman smoother (backward pass)*

Repeat Step 4 for t = T, T-1, …, 1

Step 4 – calculate $z_{t-1|T}$ and $var(z_{t-1|T})$, i.e., expectation and uncertainty about the state in period t-1, given all the data available through T

- $J_{t-1} = var(z_{t-1|t-1})C'[var(z_{t|t-1})]^{-1}$
- $z_{t-1|T} = z_{t-1|t-1} + J_{t-1}[z_{t|T} - z_{t|t-1}]$ (all data observed through T to update expectation)
- $var(z_{t-1|T}) = var(z_{t-1|t-1}) + J_{t-1}[var(z_{t|T}) - var(z_{t|t-1})]J_{t-1}'$ (all data observed through T to reduce uncertainty).

Step 5 – initialize $cov(z_{T|T}, z_{T-1|T})$, i.e., covariance of uncertainties about states at T and T-1

- $cov(z_{T|T}, z_{T-1|T}) = [I - K_TA]var(z_{t-1|t-1})$

Repeat Step 6 for t = T, T-1, …, 2

Step 6 – calculate $cov(z_{t-1|T}, z_{t-2|T})$, i.e., covariance of uncertainties about states at t-1 and t-2, given all the data available through T

- $cov(z_{t-1|T}, z_{t-2|T}) = var(z_{t-1|t-1})J_{t-2}' + J_{t-1}[cov(z_{t|T}, z_{t-1|T}) - var(z_{t-1|t-1})]J_{t-2}'$

The ultimate goal of applying the above Kalman filter and smoother is to infer, for t = 1 through

T, $z_{t|T}$, $var(z_{t|T})$ and $cov(z_{t|T}, z_{t-1|T})$, i.e., expectations, variances and lagged covariances of the state variables in any period during the observation window. Charting $z_{t|T}$'s over time would show the trend lines of the state variables, and $var(z_{t|T})$'s would determine the confidence band surrounding the trend lines.

Given the model parameters, the inferred state variables and the corresponding dynamic factor scores are determined by the complete history of the observed data, i.e., $y_t$ for t = 1 through T. When data from additional periods are observed, the above procedure needs to be applied to update the entire course of the state variables. At the end of the last observation period (T), h-step ahead forecasts can be carried out as follows, given the model parameters and $z_{T|T}$:

(A1) $\mathrm{y_{T+h|T} = A^h z_{T|T} + B}$

(A2) $\mathrm{z_{T+h|T} = C^h z_{T|T}}$.

Or, more intuitively (but equivalently),

(A3) $\mathrm{y_{T+h|T} = L f_{T+h|T} + B}$ (predicted indicator)

(A4) $\mathrm{f_{T+h|T} = \alpha_{T+h|T} + \gamma_{T+h|T}}$ (predicted dynamic factor score)

(A4.1) $\mathrm{\alpha_{T+h|T} = \alpha_{T|T} + \beta_{T|T} h + \delta_{T|T} \frac{h(h-1)}{2}}$ (predicted non-seasonal trend component)

(A4.2) $\mathrm{\beta_{T+h|T} = \beta_{T|T} + \delta_{T|T} h}$ (predicted slope of the trend)

(A4.3) $\mathrm{\delta_{T+h|T} = \delta_{T|T}}$ (predicted change rate of the slope)

(A4.4) $\mathrm{\gamma_{T+h|T} = \gamma_{T|T}}$, if $remainder(h/s) = 0$ (predicted seasonal component)

$\mathrm{\gamma_{T+h|T} = -\sum_{j=0}^{s-2} \gamma_{T-j|T}}$, if $remainder(h/s) = 1$

$\gamma_{T+h|T} = \gamma_{T-s+\mathrm{remainder}(h/s)|T}$, if $remainder(h/s) > 1$.

From the standpoint of trend projection, Equation A4.1 is of particular interest. It shows that a quadratic non-seasonal trend line can be extrapolated into the future, which has $\alpha_{T|T}$ at the origin, an initial slope of $\beta_{T|T}$ that will change at a rate of $\delta_{T|T}$ (Equation A4.2). A positive (negative) $\beta_{T|T}$ would indicate a trend line heading up (down). When $\beta_{T|T}$ and $\delta_{T|T}$ are of the same (opposite) sign, it indicates the trend will accelerate (decelerate). Given $\alpha_{T+h|T}$ and $\gamma_{T+h|T}$, we have $f_{T+h|T}$ (Equation A4), which in turn leads to predicted indicators, $y_{T+h|T}$ (Equation A3).

## Web Appendix B – Model Calibration

Although Equations 1, 2 and 2.1 through 2.4 facilitate interpretation, in order to calibrate our SDFA model it is actually more convenient to rewrite it in a state-space form:

(B1) $y_t = Az_t + B + u_t$ $\quad u_t \sim N(0, \Sigma_u)$ $\quad$ observation equation

(B2) $z_t = Cz_{t-1} + v_t$ $\quad v_t \sim N(0, \Sigma_v)$ $\quad$ state equation

In Equation B1, hereafter referred to as the *observation equation*, $y_t$ $(n \times 1)$, B $(n \times 1)$, and $u_t$ $(n \times 1)$ are the same as in Equation 1, representing, respectively, the observed indicators from period $t$, and the intercepts and the irregularities in these indicators. The irregularities are assumed to have mean zero and variance $\Sigma_u$. The difference between Equation B1 and Equation 1 lies in $z_t$, an $m \times 1$ vector of latent state variables, $z_t \stackrel{\text{def}}{=} [\alpha_t \quad \beta_t \quad \delta_t \quad \gamma_t \quad \dots \quad \gamma_{t-s-2}]'$, and $A \stackrel{\text{def}}{=} [L_{n\times p} \quad 0_{n\times 2p} \quad L_{n\times p} \quad 0_{n\times(s-2)p}]$, the matrix determining how each state variable affects each observed indicator. The dimensionality of the state variables is $m = (3 + s - 1)p$, where $s$ denotes the number of periods within a seasonal cycle, and $p$ the number of dynamic factors.

Equation B2, hereafter referred to as the *state equation*, indicates that the unobserved state variables are updated from $z_{t-1}$ to $z_t$ subject to transition matrix

$$C \overset{\text{def}}{=} \begin{pmatrix} I_p & I_p & 0 & 0 & 0 & 0 \\ 0 & I_p & I_p & 0 & 0 & 0 \\ 0 & 0 & I_p & 0 & 0 & 0 \\ 0 & 0 & 0 & -I_p & \cdots & -I_p \\ 0 & 0 & 0 & I_p & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{(s-3)p} & 0 \end{pmatrix}$$

, and shocks occurred during period $t$, i.e., $v_t \overset{\text{def}}{=} \begin{bmatrix} \varepsilon_t \\ \eta_t \\ \zeta_t \\ \xi_t \\ 0 \end{bmatrix}$, which is assumed a priori to be normally distributed with mean zero and variance

$$\Sigma_v \overset{\text{def}}{=} \begin{bmatrix} \Sigma_\varepsilon & 0 & 0 & 0 & 0 \\ 0 & \Sigma_\eta & 0 & 0 & 0 \\ 0 & 0 & \Sigma_\zeta & 0 & 0 \\ 0 & 0 & 0 & \Sigma_\xi & 0 \\ 0 & 0 & 0 & 0 & 0_{(s-2)p} \end{bmatrix}.$$

In Equations B1 and B2, i.e., the state-space form of our SDFA model, matrices $A$, B, $\Sigma_u$, and $\Sigma_v$ contain parameters to be estimated. The expected value ($\mu_0$) and uncertainty ($\Sigma_0$) of the initial state ($z_0$) also need to be estimated. We describe next the Expectation-Maximization (EM) algorithm we use to calibrate these parameters. For this description, denote the collection of all the model parameters as $\{\Theta\}$, which includes:

1. $L$, $n \times p$ matrix of loadings in Equation 1, mapping $p \times 1$ latent factors $f_t$ to $n \times 1$ observed time series $y_t$,
2. $B$, $n \times 1$ vector of intercepts in Equation 1,
3. $\Sigma_u$, $n \times n$ variance-and-covariance matrix of the idiosyncratic component ($u_t$) in Equation 1,
4. $\Sigma_\varepsilon$, $\Sigma_\eta$, $\Sigma_\zeta$ and $\Sigma_\xi$, $p \times p$ variance-and-covariance matrices in Equations 2.1 through 2.4, indicating, respectively, the sizes of stochasticity in the level ($\varepsilon_t$), slope ($\eta_t$) and slope change rate ($\zeta_t$) of the non-seasonal trend line of the latent factor, and the sizes of stochasticity in the seasonal component of the latent factor ($\xi_t$),

5. $\mu_0$ and $\Sigma_0$, mean and variance of $z_0 \stackrel{\text{def}}{=} [\alpha_0 \quad \beta_0 \quad \delta_0 \quad \gamma_0 \quad \dots \quad \gamma_{0-s-2}]'$, state of various components of the latent factors before the first observation of the manifest variables.

For parsimony, $\Sigma_u$ is assumed to be diagonal, resulting in $n$ variance terms to be estimated, as opposed to $n \times (n+1)/2$ terms. The gain in parsimony can be substantial when $n$ is large (e.g., 38 in our application). By imposing this assumption it also means that we assume all the co-movements among the observed time series are caused by the latent common factors. We also assume $\Sigma_\varepsilon$, $\Sigma_\eta$, $\Sigma_\zeta$, $\Sigma_\xi$ and $\Sigma_0$ to be diagonal because we intend to extract latent factors that are independent of one another, each capturing a distinct temporal pattern.

The model parameters $\{\Theta\}$ are to be estimated given the observed time series $y_t \in Y$ for t = 1 through T. For a given number ($p$) of latent factors, the following EM algorithm can be used for model calibration, which consists of five basic steps:

- Step 1: **Initialize the parameters** $\{\Theta^0\}$.
- Step 2: **The E-step**. Given $\{\Theta^0\}$ and Y, we use the Kalman filter (the forward pass) and Kalman smoother (the backward pass) described in Web Appendix A to derive $z_{t|T}$ and $var(z_{t|T})$. That is, given the observed data and the model parameters, we can predict the expected mean and variance of the latent factors that could have generated the observed time series.
- Step 3: **The M-step**. Given Y, by maximizing the likelihood function given $z_{t|T}$ and $var(z_{t|T})$ for t = 0 through T from Step 2, $\{\Theta^0\}$ can be updated to $\{\Theta^1\}$ as follows,

<u>For parameters related to the initial state</u>

$$\mu_0 = z_{0|T}$$

$$\Sigma_0 = \text{diag}[\text{var}(z_{0|T})]$$

<u>For parameters related to the observation equation</u>

$$A_{\alpha\gamma} \stackrel{\text{def}}{=} [I_{p\times p} \quad 0_{p\times 2p} \quad I_{p\times p} \quad 0_{p\times(s-2)p}]$$

$$z^*_{t|T} \stackrel{\text{def}}{=} A_{\alpha\gamma} z_{t|T}$$

$$[L^h \quad B^h] = \left[\sum_{t=1}^{T} y_t^h z_{t|T}^{*}{}' \quad \sum_{t=1}^{T} y_t^h\right]\begin{bmatrix} \sum_{t=1}^{T} z_{t|T}^{*} z_{t|T}^{*}{}' + A_{\alpha\gamma} var(z_{t|T}) A_{\alpha\gamma}{}' & \sum_{t=1}^{T} z_{t|T}^{*} \\ \sum_{t=1}^{T} z_{t|T}^{*}{}' & T \end{bmatrix}^{-1}$$

$\Sigma_u^h = \frac{1}{T}\sum_{t=1}^{\mathrm{T}}\left[(y_t^h - L^h z_{t|T}^{*} - B^h)(y_t^h - L^h z_{t|T}^{*} - B^h)' + (L^h A_{\alpha\gamma}) var(z_{t|T})(L^h A_{\alpha\gamma})'\right]$, where the superscript $h$ indicates the $h$-th row of the corresponding matrix.

For parameters related to the state equation

$\Sigma_\varepsilon^{\alpha_j} = \frac{1}{T}\sum_{t=1}^{\mathrm{T}}\left[\left(z_{t|T}^{\alpha_j} - C^{\alpha_j} z_{t-1|T}\right)\left(z_{t|T}^{\alpha_j} - C^{\alpha_j} z_{t-1|T}\right)' + I^{\alpha_j} var(z_{t|T}) I^{\alpha_j}{}' + C^{\alpha_j} var(z_{t-1|T}) C^{\alpha_j}{}' - I^{\alpha_j} cov(z_{t|T}, z_{t-1|T}) C^{\alpha_j}{}' - C^{\alpha_j} cov(z_{t|T}, z_{t-1|T}) I^{\alpha_j}{}'\right]$, where the superscript $\alpha_j$ indicates the row corresponding to the $\alpha$ component of the $j$-th factor.

Formula of the same form can be applied to calculate $\Sigma_\eta^{\beta_j}$, $\Sigma_\zeta^{\delta_j}$ and $\Sigma_\xi^{\gamma_j}$

- Step 4: **Go back to Step 2** with updated $\{\Theta^1\}$ and iterate until convergence.
- Step 5: **Varimax Rotation**. Like standard factor analysis or any other factor-analytic model, our SDFA model is invariant to orthogonal rotation. In other words, any orthogonal rotation (i.e., $Q$ s.t. $Q \times Q' = I$) to the loadings matrix ($L$) along with an inverse rotation to the factor scores ($f$) would produce the same fit to the observed time series (i.e., $L \times f = (L \times Q) \times (Q' \times f) \stackrel{\text{def}}{=} L^* \times f^*$). To obtain a unique solution for the factor loadings and scores, we apply a Varimax rotation to $L$ and $z_{t|T}$, seeking a distinctive factor structure such that, for each factor, large loadings will result for a few variables and the rest will be small[2].

Lastly, in our empirical implementation we standardized all the observed time series before applying the EM algorithm given above. It is important to note that standardization will not lead to different trends. To see this, let $y_t^* = \Sigma^{-1} y_t$, where $y_t^*$ is the standardized series and $\Sigma$ is a diagonal

[2] Varimax, the most common factor rotation routine in marketing research, should be available in most popular software packages (e.g., Proc Factor in SAS). For more on Varimax and other common factor rotation routines (e.g., Quartimax and Eqimax), see Basilevsky (1994), *Statistical Factor Analysis and Related Method: Theory and Applications*. John Wiley & Sons; and Kaiser (1958), “The Varimax Criterion for Analytic Rotation in Factor Analysis,” *Psychometrika*, 23 (3).

matrix containing standard deviations. As a result, the observation equation (Equation 1, $y_t = Lf_t + B + u_t$) is equivalent to $y_t^* = L^* f_t + B^* + u_t^*$, where $L^* = \Sigma^{-1}L$, $B^* = \Sigma^{-1}B$, $u_t^* = \Sigma^{-1}u_t$. This shows that $f_t$, the latent factor scores, will not be affected at all and the effect of standardization is equivalent to dividing the factor loadings, intercept and noise term by the corresponding time series' standard deviation. After standardizing each time series, we can interpret the factor loading estimates ($\hat{L}$) as "for every unit change in the latent factor, one would expect to see a change of $\hat{L}$ standard deviations in the corresponding time series." Without standardization, differences in factor loading estimates would confound differences in the scale/variance of the times series with differences in the correlation between the latent factor and the series. In analyses such as ours, one care about correlation between the factors and the observed indicators, as opposed to their covariance.

## Web Appendix C – Split-Half Reliability Test

Our SDFA model assumes that the latent factor loadings ($L$) will be time-invariant during the observation window, and any observed temporal variations in the time series are caused by variations in the latent factor scores. One might question such an assumption and wonder whether the structure of co-movements among the time series may change over time as well. To address this concern, we conducted a split-half reliability test, where we recalibrated our model twice, once with data from the first half of the observation window and another with data from the second half. The table below reports three sets of correlations between the estimated loadings: 1) the full sample and the first half, 2) the full sample and the second half, and 3) the first half and the second half. Strong correlation coefficients in all cases indicate that the factor loadings structure has remained largely stable across the samples (full vs. first vs. second half).

| Factor \ Correlation | Full sample & first half | Full sample & second half | First half & second half |
|---|---|---|---|
| Foreign Mass | 0.92 | 0.93 | 0.95 |
| U.S. Mass | 0.83 | 0.92 | 0.89 |
| Euro Lux | 0.92 | 0.89 | 0.92 |
| GM Survived | 0.92 | 0.93 | 0.86 |
| Lexus / (RAM) | 0.88 | 0.74 | 0.76 |
| Subaru / (Mazda & Saturn) | 0.66 | 0.82 | 0.69 |
| GM Cancelled & Isuzu / (Hyundai Kia & Suzuki) | 0.74 | 0.87 | 0.81 |

The above results aside, we stress the need to distinguish between non-stationarity in the trend lines and non-stationarity in the loadings. Major shifts can take place in the trend lines with no significant changes in the loadings structure, which is determined by co-movements among the observed time series. Regardless of the values of the observed indicators, as long as they follow the same co-movement patterns, the resulting loadings will remain unchanged. To see this, suppose there was a sudden increase in gas prices due to a crisis in the Middle-East, and as a result the trend line representing consumer interest in fuel-efficiency spiked. Despite this sudden change, as long as vehicles of similar fuel efficiency are affected in a similar way, the correlation between them would remain the same, which would result in the same loadings structure.