Modeling Ideas: Transit Modeling for Kepler Light Curves

Paul Baines

August 21, 2013

Disclaimer: These are just preliminary ideas. There are likely to be typos etc. throughout the document. Use at your own risk. ☺

1 Modeling Simple Aperture Photometry Kepler Light Curves

Consider a random process Y defined as a function of time and depending on unknown parameter(s) θ i.e., $Y(t|\theta)$. For simplicity let $\theta = (\eta, \psi, \xi)$ and $\mathcal{T}$ be the interval on which the process is defined. We assume that:

$$\begin{aligned} Y_{sap}(t) &= m_{sap}(t|\phi, \eta, \xi) + \epsilon_{sap}(t|\psi), \qquad \forall\, t \in \mathcal{T} \\ m_{sap}(t|\phi, \eta, \xi) &= (\mu(t|\phi) + a(t|\xi))(1 - q(t|\eta)) \end{aligned}$$

where $\epsilon(t|\psi)$ is a mean-zero random process, $\mu(t|\phi)$ is an overall mean corresponding to low-frequency image artifacts (i.e., the objects to be removed by detrending), $a(t|\xi)$ is a random process designed to capture any remaining non-transit-related signals (probably astrophysical) remaining after detrending, and $q(t|\eta)$ is the multiplicate reduction in the expected intensity due to the transit i.e., $q(t|\eta) \equiv 0$ if there are no transits. In the event of a transit we observe a version of $m(t|\xi)$ 'dampened' by a multiplicative amount $q(t|\eta)$. Typically, for detrended data the component $\mu(t|\phi)$ is divided out to give:

$$\begin{aligned} Y_d(t|\phi) = \frac{Y_{sap}(t)}{\mu(t|\phi)} &= m_d(t|\phi, \eta, \xi) + \epsilon_d(t|\phi, \psi), \qquad \forall\, t \in \mathcal{T} \qquad (1) \\ m_d(t|\phi, \eta, \xi) &= \left(1 + \frac{a(t|\xi)}{\mu(t|\phi)}\right)(1 - q(t|\eta)) \equiv (1 + f(t|\phi, \xi))(1 - q(t|\eta)), \\ \epsilon_d(t|\phi, \psi) &\equiv \frac{\epsilon_{sap}(t|\psi)}{\mu(t|\phi)}. \end{aligned}$$

In this detrended setting the interpretation of q remains the same, and $f(t|\phi, \xi)$ models the astrophysical processes (and other artefacts not removed by detrending).

2 Modeling Detrended Kepler Light Curves

For simplicity we now focus on fitting detrended light curves of the form shown in (1) and drop the subscripts for notational brevity. Our model for this section will be:

$$Y(t|\phi) = (1 + f(t|\phi, \xi))(1 - q(t|\eta)) + \epsilon(t|\phi, \psi), \qquad \forall\, t \in \mathcal{T}. \qquad (2)$$

There is a trade-off in the complexity we place in f and in ϵ, but for our purposes we assume a mean-zero Gaussian process for ϵ:

$$\epsilon(t) \sim \mathrm{GP}\left(0, \Sigma_\epsilon(\psi)\right), \tag{3}$$

where $\Sigma_\epsilon(t_i, t_j|\psi) = K(|t_i - t_j|; \psi)$ i.e., stationary covariance. Although the form in (3) is fairly rich, in practice we will usually assume either independent noise with constant variance, or independent noise with different variances for different segments of the time series to reflect phenomena such as CCD changes i.e.,

$$\Sigma_\epsilon(t_i, t_j) = \sigma^2_{\epsilon,q(t_i)}\delta_{t_i t_j}, \qquad q(t_i) \in \{1, \ldots, n_s\},$$

where $q(t_i)$ is an increasing sequence corresponding to the index of the segment in which time point i falls and n_s denotes the number of 'segments' in the full time series. In this case $\psi = \{\sigma^2_{\epsilon,1}, \ldots, \sigma^2_{\epsilon,n_s}\}$. For reasons we will see shortly, we often we take n_s to be the number of quarters in the series, denoted by R i.e., assuming a constant variance within each quarter, but allowing variances to differ by quarter (or 'segment'). We assume here that the change-points for the noise process are known and specified by the analyst.

Although we only observe a noisy version of $(1+f)(1-q)$, the goal is to 'separate' this underlying mean into two pieces: one representing 'transit-like variation', the other representing 'everything else'. Fortunately we are able to approximate the form of q parametrically. Let t_0 denote the time of the first transit, t_d the transit duration, α the multiplicative reduction in the signal and P the period. Then we assume:

$$q(t|\eta) = \begin{cases} 0 & \text{if } (t \bmod P) \notin [t_0, t_0 + t_d] \\ \alpha & \text{if } (t \bmod P) \in [t_0, t_0 + t_d] \end{cases},$$

where $\eta = \{t_0, t_D, \alpha, P\}$. More sophisticated models for the transit could also be used to allow for smooth transits. The remaining modeling task is to decide upon a structure for f. The complexity of f and ϵ is heavily dependent on the detrending process, and thus we would prefer to jointly model the detrending and transit signal, something we discuss in section 4. Even with aggressive detrending we still would like to have flexible models are required to capture all possible remaining instrumental and astrophysical signals. To try to achieve this, we decide to model f in the wavelet domain. For computational tractability we apply wavelet transforms to each quarter of data separately. Let y_r, f_r and q_r denote subvectors of y, f and q that correspond to quarter $r = 1, \ldots, R$.

$$\begin{aligned} Y_r(t) &= (1 + f_r(t|\xi))(1 - q_r(t|\eta)) + \epsilon_r(t), \\ \Rightarrow \quad (I - Q_r)^{-1} y_r - 1 &= f_r + (I - Q_r)^{-1}\epsilon_r. \end{aligned} \tag{4}$$

Taking the DWT of both sides yields:

$$\tilde{y}_r(q_r) = w_r + \tilde{\epsilon}_r, \tag{5}$$

where $\tilde{y}_r(q_r) = \mathcal{W}_r\left[(I - Q_r)^{-1} y_r - 1\right]$, $w_r = \mathcal{W} f_r$ and $\tilde{\epsilon}_r = \mathcal{W}(I - Q_r)^{-1}\epsilon_r$, so:

$$\begin{aligned} \tilde{\epsilon}_r &\sim N(0, \Sigma_{\tilde{\epsilon}_r}), \\ \Sigma_{\tilde{\epsilon}_r} &= \mathcal{W}(I - Q_r)^{-1}\mathrm{Var}(\epsilon_r)(I - Q_r)^{-1}\mathcal{W}^T. \end{aligned}$$

If we assume that $\mathrm{Var}(\epsilon_r) = \sigma^2_{\epsilon,r} I$ then $\Sigma_{\tilde{\epsilon}_r}$ is not a diagonal matrix, potentially leading to a large computational burden. Note that even if the transit signal only occurs in a small fraction of the

time series so that only a small portion of the diagonal elements are different from one, $\Sigma_{\tilde{\epsilon}}$ will still be a dense matrix. More importantly, we note that the covariance matrix depends on the transit model, and thus if we wish to assume that $\mathrm{Var}(\epsilon_r)$ is such that $\Sigma_{\tilde{\epsilon}_r}$ is diagonal then $\mathrm{Var}(\epsilon_r)$ will need to depend on the transit model. If we assume that:

$$\mathrm{Var}(\epsilon_r) = \sigma^2_{\tilde{\epsilon},r}(I - Q_r)^2, \tag{6}$$

then we obtain that:

$$\Sigma_{\tilde{\epsilon}_r} = \sigma^2_{\tilde{\epsilon},r} I, \qquad \Sigma_{\tilde{\epsilon}} = \text{block-diag}\left(\Sigma_{\tilde{\epsilon}_1}, \ldots, \Sigma_{\tilde{\epsilon}_R}\right).$$

Note that in the current formulation, the assumption in (6) necessitates that 'segments' of the series for which the variances are constant correspond to the 'segments' for which the Wavelet transforms are taken (typically quarters). To allow for discontinuities in the variance process that do not match with the segments corresponding to the wavelet transform leads to computational challenges (i.e., large non-diagnonal covariance matrices) and is best approached by performing all modeling directly in the time domain. However, while we require that the variance is constant within wavelet-transformed chunks of the time series, it is worth noting that the length of the chunks can be different (leading to different numbers of wavelet coefficients for each chunk). For non-power-of-two chunk lengths, standard wavelet techniques (e.g., mirroring, padding) must be used to allow for use of the DWT. For notational simplicity we also denote:

$$\Sigma_{\tilde{\epsilon}} = \mathrm{diag}(X\vec{\sigma}_{\tilde{\epsilon}}), \tag{7}$$

where X is an $n_w \times R$ matrix, with a single 1 per row, in the column location corresponding to the segment that the point in that row belongs to. Here n_w is the number of stacked wavelet coefficients for the time series, which may or may not match n, the number of time points, depending on the padding/mirroring strategy used. As expected, $\vec{\sigma}_{\tilde{\epsilon}} = (\sigma^2_{\tilde{\epsilon},1}, \sigma^2_{\tilde{\epsilon},2}, \ldots, \sigma^2_{\tilde{\epsilon},R})^T$ is vector of the unique noise variance parameters. Within each chunk, the plausibility of the assumption in (6) should be reasonable in most cases since $(1 - Q)^2$ is very close to the identity matrix, but implying a small reduction in the noise variance duration any transits. For Poisson-type noise this is expected since the fractional reduction in the mean during transits implies a corresponding fractional reduction in the noise variance during transit. In either case, we expect the practical implication of this assumption to be relatively small. We then have:

$$\tilde{y}_r(q_r)|\eta, \sigma^2 \sim N\left(w_r, \sigma^2_r I\right), \qquad r = 1, \ldots, R.$$

To pool information within the wavelet coefficients of each quarter and across quarters we assume a hierarchical structure for the w_r. Let:

$$w_{1:R} = \begin{pmatrix} w_1 \\ \vdots \\ w_R \end{pmatrix},$$

denote the stacked vector of all wavelet coefficients then we assume that:

$$w_{1:R}|d, \Sigma_w \sim N\left(Ad, \Sigma_{w_{1:R}}\right), \tag{8}$$

where A is a known $n_w \times n_a$ matrix that allows for pooling of means of (linear combinations of) wavelet coefficients within wavelet scales and across time periods. For computational simplicity, given the large dimensionality of $\Sigma_{w_{1:R}}$, we assume it to be a diagonal, or otherwise rapidly

invertible (e.g., block-diagonal, Toeplitz etc.), matrix. One method of achieving this is to let $\Sigma_{w_{1:R}} = \text{diag}(B\vec{\sigma}_w)$ where B is a $n_w \times n_b$ matrix that specifies the pooling of the covariance parameters within and across periods and wavelet scales and $\vec{\sigma}_w = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_{n_b}^2)^T$ is a vector of the unique variance parameters. If the same pooling is selected for both means and covariances then $A = B$. To make things concrete, in the extreme case, if $n_a = n_w$ and A is the identity matrix then d is an $n_w \times 1$ vector, and the full model contains $n_w + p$ parameters where p is the number of additional parameters. This massively overparametrized model has no pooling and thus illustrates the need for pooling. Also note that n_w need not match n, the number of time points in the series, depending on the mirroring/padding process used in computing the wavelet transforms for non-power-of-two lengths.

In the other extreme, we could consider $n_a = 1$ and set $A = (1, 0, 0, \ldots, 0, 1, 0, \ldots)^T$ with ones corresponding to the scaling coefficients and zeroes corresponding to the detail coefficients. In this formulation, the mean of the wavelet coefficients for all scaling coeffiicents is set to d, which is constant across periods. The detail wavelet coefficients are then given mean zero for all levels and all time periods. This version drastically reduces the complexity of the model by applying a highly structured model on the wavelet coefficients, but imposes undesirable features such as identical mean structures for all periods.

For an intermediate case, consider $n_a = n_{w,q}$ where $n_{w,q}$ is the number of wavelet coefficients for each quarter of data. In this case, we could set A to be of the form:

$$A = \begin{pmatrix} I_{n_{w,q}} \\ \vdots \\ I_{n_{w,q}} \end{pmatrix},$$

thus successive quarters essentially provide replicates for each wavelet coeffiicent. Such a model provides across-period pooling, but no pooling within wavelet scales which may be desirable.

To achieve this feature, consider the special case that $n_a = J$, the number of wavelet levels. If A contains zeros and at most a single one per row, then Ad is a vector that contains one of the d_j's in each position (or a zero), achieving pooling across times and within wavelet scales.

For the covariance, recall that we would like $\Sigma_{w_{1:R}}$ to be diagonal. As mentioned above, one intuitive way to achieve this is to let $\Sigma_{w_{1:R}} = \text{diag}(B\vec{\sigma}_w)$ where B has exactly one 1 entry per row, and $\vec{\sigma}_w$ is a vector of the unique variance parameters.

Letting $\xi = \{d, \Sigma_w\}$ where $d = \{d_1, \ldots, d_{n_a}\}$ and $\Sigma_w = \left\{\sigma_1^2, \ldots, \sigma_{n_b}^2\right\}$, and $\psi = \left\{\sigma_1^2, \ldots, \sigma_R^2\right\}$, the parameters of the model are therefore $\theta = \{\psi, \eta, \xi\}$. With 4 transit parameters, the model contains a total of $4 + R + n_a + n_b$.

Note that due to the nature of the wavelet basis we expect the mean of wavelet coefficients to be zero (or close to) for all of the details and potentially non-zero for the scaling coefficients. Therefore, it may be feasible to reduce the number of parameters in d by a large amount. In all cases we assume conjugate priors for d of the form:

$$d \sim N(m_{d,0}, V_{d,0}), \tag{9}$$

Combining (5) and (8) we obtain:

$$\begin{aligned} \tilde{y}_{1:R}(q_{1:R})|\psi, \eta, \xi &\sim N\left(Ad, U_{1:R}^{-1}(\psi, \Sigma_w)\right), \\ U_{1:R}(\psi, \Sigma_w) &= \left[\Sigma_{\tilde{\epsilon}} + \Sigma_{w_{1:R}}\right]^{-1}, \end{aligned} \tag{10}$$

which allows us to sample from the conditional posterior of all elements of d directly, without conditioning on w. In the special case of $\Sigma_{w_{1:R}} = \text{diag}(B\vec{\sigma}_w)$ and $\Sigma_{\tilde{\epsilon}} = \text{diag}(X\vec{\sigma}_{\tilde{\epsilon}})$ we obtain:

$$U_{1:R}(\psi, \Sigma_w) = \left[\text{diag}(X\vec{\sigma}_{\tilde{\epsilon}}) + \text{diag}(B\vec{\sigma}_w)\right]^{-1}. \tag{11}$$

This partial marginalization is analogous to methods for sampling in normal-normal hierarchical models. For example, from the marginal model in (10) we get:

$$d|\eta, \psi, \Sigma_w \sim N(m_{d,n}, V_{d,n}),$$

where:

$$V_{d,n} = \left[A^T U_{1:r}(\psi, \Sigma_w) A + V_{d,0}^{-1}\right]^{-1},$$
$$m_{d,n} = V_{d,n}\left[A^T U_{1:r}(\psi, \Sigma_w)\tilde{y}_{1:r}(q_r) + V_{d,0}^{-1} m_{d,0}\right].$$

Though the above form looks computationally intensive, recall that the forms of Σ_w and A can easily be chosen so that $A^T U_{1:r}(\psi, \Sigma_w) A$ is diagonal. Even if this is not the case, the dimensionality of the matrices in this step are just $(t_d \times t_d)$ and t_d will typically be chosen to be no more than a few hundred.

Similarly to the marginalization over w, we can integrate out d from (10) to obtain:

$$\tilde{y}_{1:R}(q_{1:R})|\psi, \eta, \Sigma_{w_{1:R}}, \sim N\left(Am_{d,0}, Z_{1:R}^{-1}(\psi, \Sigma_{w_{1:R}})\right), \tag{12}$$
$$Z_{1:R}(\psi, \Sigma_w) = \left[\Sigma_{\tilde{\epsilon}} + \Sigma_{w_{1:R}} + AV_{d,0}A^T\right]^{-1}.$$

which allows us to sample from the conditional posterior of all elements of $(\psi, \eta, \Sigma_{w_{1:R}})$ directly, without conditioning on w (although in practice this may not be advisable). Again, the special case mentioned earlier simplifies to:

$$Z_{1:R}(\psi, \Sigma_w) = \left[\text{diag}(X\vec{\sigma}_{\tilde{\epsilon}}) + \text{diag}(B\vec{\sigma}_w) + AV_{d,0}A^T\right]^{-1}. \tag{13}$$

However, it is important to note that $AV_{d,0}A^T$ is typically low-rank and non-diagonal (since $n_a << n_w$), so this identity can be less useful in practice.

2.1 Handling Missing Data

The model outlined in the previous sections can be extended to handle missing data in the time domain. Let $t_{mis} = \{t_{mis,1}, \ldots, t_{mis,n_{mis}}\}$ denote the set of time points for which the value of $Y(t)$ was not observed, similarly for t_{obs}. We introduce some new notation:

$$y_{com} = y, \qquad y_{mis} = \begin{pmatrix} y(t_{mis,1}) \\ y(t_{mis,2}) \\ \cdots \\ y(t_{mis,n_{mis}}) \end{pmatrix}, \qquad y_{obs} = \begin{pmatrix} y(t_{obs,1}) \\ y(t_{obs,2}) \\ \cdots \\ y(t_{obs,n_{mis}}) \end{pmatrix}.$$

Here Y_{mis} becomes another latent variable to be integrated out in the posterior distribution. Relative to cases with complete data, the only additional sampling step is to sample y_{mis} from its conditional posterior distribution. Fortunately, by standard theory, Y_{mis} is seen to be normally distributed. Let $\Sigma_{\epsilon,oo}(\psi)$ and $\Sigma_{\epsilon,mm}(\psi)$ denote the submatrices of the covariance matrix $\Sigma_\epsilon(\psi)$ that correspond to the observed and missing matrices respectively. Similarly, let $\Sigma_{\epsilon,om} = \Sigma_{\epsilon,mo}^T$ denote the covariance matrix between the observed and missing portions of the data. We again work from the partially marginalized posterior in (10):

$$\tilde{y}_{1:R}(q_{1:R})|\psi, \eta \sim N\left(Ad, U_{1:R}^{-1}(\psi, \Sigma_w)\right). \tag{14}$$

Let:

$$\mu_{obs} = (I - Q_{obs})f_{obs}, \qquad \mu_{mis} = (I - Q_{mis})f_{mis},$$

where Q_{obs}, Q_{mis}, f_{obs} and f_{mis} are the subcomponents of Q and f corresponding to the observed and missing components respectively. The conditional distribution of Y_{mis} can then be seen to be:

$$Y_{mis}|y_{obs}, \psi, f \sim N\left(\mu_{mis} + \Sigma_{\epsilon,mo}\Sigma_{\epsilon,}^{-1}(y_{obs} - \mu_{obs}), \Sigma_{\epsilon,mm} - \Sigma_{\epsilon,mo}\Sigma_{\epsilon,oo}^{-1}\Sigma_{\epsilon,om}\right).$$

2.2 Computational Framework

For the class of models presented in section 1, we now describe the general computational framework. Recall that the parameters of our model are $\theta = \{\psi, \eta, d, \Sigma_w\}$. For clarity we restate the model and prior assumptions as well as some useful marginalizations:

1. Model assumptions (Blockwise):

$$\begin{aligned}
\tilde{y}_r(q_r) &= \mathcal{W}_r\left[(I - Q_r)^{-1}y_r - 1\right] \\
\tilde{y}_r(q_r)|\eta, \sigma_r^2 &\sim N\left(w_r, \sigma_{\tilde{\epsilon},r}^2 I\right), \qquad r = 1, \ldots, R, \\
w_r|d, \Sigma_w &\sim N\left(A_{[r]}d, (\Sigma_{w_{1:R}})_{[r,r]}\right),
\end{aligned}$$

   where the subscripts $[r]$ and $[r, r]$ denoted the sub-blocks corresponding to segment r.

2. Model assumptions (Stacked):

$$\begin{aligned}
\tilde{y}_{1:R}(q_{1:R})|\eta, \psi &\sim N\left(w_{1:R}, \Sigma_{\tilde{\epsilon}}\right), \\
w_{1:R}|d, \Sigma_w &\sim N\left(Ad, \Sigma_{w_{1:R}}\right).
\end{aligned}$$

   Special case:

$$\begin{aligned}
\tilde{y}_{1:R}(q_{1:R})|\eta, \psi &\sim N\left(w_{1:R}, \mathrm{diag}(X\vec{\sigma}_{\tilde{\epsilon}})\right), \\
w_{1:R}|d, \Sigma_w &\sim N\left(Ad, \mathrm{diag}(B\vec{\sigma}_w)\right).
\end{aligned}$$

3. Priors:

$$\begin{aligned}
p(\psi): &\quad \sigma_r^2 \sim \text{Inv-}\chi^2(\nu_{r,0}, s_{r,0}^2), \quad r = 1, \ldots, R, \\
p(\eta): &\quad p(t_0, t_d, P, \alpha) = p(\alpha)p(P)p(t_d|P)p(t_0|t_d, P) \\
p(d): &\quad d \sim N(m_{d,0}, V_{d,0}), \\
p(\Sigma_w): &\quad \sigma_{w,j}^2 \sim \text{Inv-}\chi^2(\nu_{w,j,0}, s_{w,j,0}^2), \quad j = 1, \ldots, n_b.
\end{aligned}$$

4. Partial marginalizations:

   - Integrating out $w_{1:r}$ gives the prior form:

$$\begin{aligned}
\tilde{y}_{1:R}(q_{1:R})|\psi, \eta, d, \Sigma_w &\sim N\left(Ad, U_{1:R}^{-1}(\psi, \Sigma_w)\right), \\
U_{1:R}(\psi, \Sigma_w) &= \left[\Sigma_{\tilde{\epsilon}} + \Sigma_{w_{1:R}}\right]^{-1}, \\
\text{(Special case):} \qquad U_{1:R}(\psi, \Sigma_w) &= \left[\mathrm{diag}(X\vec{\sigma}_{\tilde{\epsilon}}) + \mathrm{diag}(B\vec{\sigma}_w)\right]^{-1}.
\end{aligned}$$

The posterior for d from this marginalized form is then:

$$d|\psi, \eta, \Sigma_w \sim N(m_{d,n}, V_{d,n}),$$

where:

$$V_{d,n} = \left[A^T U_{1:R}(\psi, \Sigma_w) A + V_{d,0}^{-1}\right]^{-1},$$
$$m_{d,n} = V_{d,n}\left[A^T U_{1:R}(\psi, \Sigma_w)\tilde{y}_{1:R}(q_r) + V_{d,0}^{-1} m_{d,0}\right].$$

- Integrating out $w_{1:r}$ and d:

$$\tilde{y}_{1:R}(q_{1:R})|\psi, \eta, \Sigma_w, \sim N\left(Am_{d,0}, Z_{1:R}^{-1}(\psi, \Sigma_w)\right),$$
$$Z_{1:R}(\psi, \Sigma_w) = \left[\Sigma_{\tilde{\epsilon}} + \Sigma_{w_{1:R}} + AV_{d,0}A^T\right]^{-1},$$
$$\text{(Special case):} \qquad Z_{1:R}(\psi, \Sigma_w) = \left[\text{diag}(X\vec{\sigma}_{\tilde{\epsilon}}) + \text{diag}(B\vec{\sigma}_w) + AV_{d,0}A^T\right]^{-1}.$$

2.3 Computational Details

The likelihood for this version of the model is given by:

$$p(y, w|\theta) \propto p(y|w, \eta, \psi)p(w|d, \Sigma_w), \tag{15}$$
$$\Rightarrow \qquad p(y|\theta) = \int p(y, w|\theta)dw.$$

The full Bayesian version simply requires a prior on all components of θ and may be preferable,

$$p(\theta, w|y) \propto p(y|w, \eta, \psi)p(w|d, \Sigma_w)p(\psi, \eta, d, \Sigma_w), \tag{16}$$
$$\Rightarrow \qquad p(\theta|y) = \int p(w, \theta|y)df,$$

The form of (16) lends itself to a natural Gibbs sampler of the form:

$$[\psi|w, \eta, y], \quad \left[\eta|\sigma^2, w\right], \quad [w|\eta, \psi, d, \Sigma_w], \quad [d|\eta, \psi, w, \Sigma_w], \quad [\Sigma_w|d, w].$$

Sampling ψ: With a conjugate prior on $\sigma^2_{\tilde{\epsilon},r}$ of the form $\sigma^2_{\tilde{\epsilon},r} \sim \text{Inv-}\chi^2(\nu_{r,0}, s^2_{r,0})$ we obtain:

$$\sigma^2_{\tilde{\epsilon},r}|d, \eta, \Sigma_w \sim \text{Inv-}\chi^2(\nu_{r,n}, s^2_{r,n}), \qquad r = 1, \ldots, R,$$

where

$$\nu_{r,n} = \nu_{r,0} + n_{d,r},$$
$$s^2_{r,n} = \frac{\nu_{r,0}s^2_{r,0} + (\tilde{y}_r(q_r) - w_r)^T(\tilde{y}_r(q_r) - w_r)}{\nu_{r,0} + n_{d,r}},$$

and $n_{d,r}$ is the number of points in quarter r (recall that this can vary if segments are selected to be of different lengths).

Sampling η: Sampling for the parameters controlling the transit cannot be done in closed form. However, from (5), we see that values of η that produce transits that match features of $\tilde{y}_r(f_r)$ will be given higher posterior probabilities. Note that this accounts for data across all quarters, while also allowing for different noise variances for each quarter of Kepler observations. Simple box-least squares (BLS) algorithms can be used to generate a proposal distribution for η, or random walk proposals can be used if the chain is in the neighborhood of non-negligible posterior density. More efficient algorithms for exploring the parameter space may be available, but we defer discussion of these until testing on real data.

Using MH-type updates the log-posterior terms reduce to:

$$\log p(\eta) - \frac{1}{2}(\tilde{y}_{1:R}(q_{1:R}) - w_{1:R})^T \Sigma_{\tilde{\epsilon}}^{-1} (\tilde{y}_{1:R}(q_{1:R}) - w_{1:R}) = \log p(\eta) - \frac{1}{2} \sum_{r,j} \frac{(\tilde{y}_{r,j}(q_r) - w_{r,j})^2}{\sigma_r^2}$$

More sophisticated energy-based or HMC-based moves can also be used in place of a vanilla MH-update.

Sampling w: Recall that we have:

$$\begin{aligned} \tilde{y}_{1:R}(q_{1:R}) | \eta, \psi &\sim N\left(w_{1:R}, \Sigma_{\tilde{\epsilon}}\right), \\ w_{1:R} | d, \Sigma_w &\sim N\left(Ad, \Sigma_{w_{1:R}}\right), \end{aligned}$$

so:

$$w_{1:R} | \tilde{y}_{1:R}(q_{1:R}), \psi, \Sigma_{w_{1:R}} \sim N\left(m_{w,1:R}, V_{w,1:R}\right), \tag{17}$$

where:

$$\begin{aligned} V_{w,1:R} &= \left[\Sigma_{\tilde{\epsilon}}^{-1} + \Sigma_{w_{1:R}}^{-1}\right]^{-1}, \\ m_{w,1:R} &= V_{w,1:R}\left[\Sigma_{\tilde{\epsilon}}^{-1} \tilde{y}_{1:R}(q_{1:R}) + \Sigma_{w_{1:R}}^{-1} Ad\right]. \end{aligned}$$

Recall that $\Sigma_{\tilde{\epsilon}}$ is diagonal and $\Sigma_{w_{1:R}}$ is either assumed to be diagonal or otherwise rapidly invertible, thus the sampling in (17) can be performed efficiently as long as a Cholesky (or other matrix root) of $V_{w,1:R}$ can be found rapidly. In the special case we obtain:

$$\begin{aligned} V_{w,1:R} &= \left[\mathrm{diag}(X\vec{\sigma}_{\tilde{\epsilon}})^{-1} + \mathrm{diag}(B\vec{\sigma}_w)^{-1}\right]^{-1}, \\ m_{w,1:R} &= V_{w,1:R}\left[\mathrm{diag}(X\vec{\sigma}_{\tilde{\epsilon}})^{-1} \tilde{y}_{1:R}(q_{1:R}) + \mathrm{diag}(B\vec{\sigma}_w)^{-1} Ad\right]. \end{aligned}$$

Sampling d: Assuming a conjugate prior for d:

$$d \sim N(m_{d,0}, V_{d,0}),$$

then, from the marginal model in (10) we get:

$$d | w, \Sigma_w \sim N(m_{d,n}, V_{d,n}),$$

where:

$$\begin{aligned} V_{d,n} &= \left[A^T U_{1:R}(\psi, \Sigma_w) A + V_{d,0}^{-1}\right]^{-1}, \\ m_{d,n} &= V_{d,n}\left[A^T U_{1:R}(\psi, \Sigma_w) \tilde{y}_{1:R}(q_{1:R}) + V_{d,0}^{-1} m_{d,0}\right]. \end{aligned}$$

Recall that:

$$U_{1:R}(\psi, \Sigma_w) = \left[\Sigma_{\tilde{\epsilon}} + \Sigma_{w_{1:R}}\right]^{-1},$$
$$\text{(Special case):} \qquad U_{1:R}(\psi, \Sigma_w) = \left[\text{diag}(X\vec{\sigma}_{\tilde{\epsilon}}) + \text{diag}(B\vec{\sigma}_w)\right]^{-1}.$$

so again, it can be seen that if $U_{1:R}$ is typically diagonal or otherwise rapidly invertible, thus for the form of A discussed earlier, $V_{d,n}$ is typically also diagonal, so the calculations can be done very rapidly.

Sampling $\Sigma_{w_{1:R}}$: Assuming that $\Sigma_{w_{1:R}} = \text{diag}(\sigma_1^2, \sigma_1^2, \ldots, \sigma_{n_b}^2)$ (i.e., possible duplicates on the diagonal), and placing a conjugate prior on $\sigma_{w,j}^2$ of the form $\sigma_{w,j}^2 \sim \text{Inv-}\chi^2(\nu_{w,j,0}, s_{w,j,0}^2)$ we can obtain:

$$\sigma_{w,j}^2 \sim \text{Inv-}\chi^2(\nu_{w,j,n}, s_{w,j,n}^2),$$

where $\nu_{w,j,n}$ and $s_{w,j,n}^2$ are posterior parameters corresponding to the choice of B matrix (the notation is more involved, but these are straightforward to compute). Recall that:

$$w_{1:R}|d, \Sigma_w \sim N\left(Ad, \text{diag}(B\vec{\sigma}_w)\right),$$

where $\Sigma_w = (\sigma_1^2, \ldots, \sigma_{n_b}^2)$ i.e., Σ_w contains unique elements only, with no duplicates on the diagonal. Recall that B is an $n_w \times n_b$ matrix, where n_w is the total number of stacked wavelet coefficients (depending on padding, mirroring etc., this may not exactly match n, the number of time-points). Then:

$$\nu_{w,j,n} = \nu_{w,j,0} + n_j,$$
$$s_{w,j,n}^2 = \frac{\nu_{w,j,0} s_{w,j,0}^2 + (w_{1:R} - Ad)_{[j]}^T (w_{1:R} - Ad)_{[j]}}{\nu_{r,0} + n_j},$$

where n_j is the number of points which have variance σ_j^2 (i..e, the column sum of column j of B) and the subscript $[j]$ corresponds to the rows that have a 1 in column j.

Carry on from here...

3 Specific Examples

Given the framework in section 1 is very general, we now present some concrete examples of models within this framework.

3.1 Pooling Information Within Wavelet Scales

3.2 Pooling Information Across Quarters

3.3 Wavelet-based Joint Pooling Across Time and Scale

3.4 Other Comments

Σ_w trickier. What about non-diagonal? Should it really be $A\Sigma_w A^T$? Or it could be a different matrix i.e., $B\Sigma_w B^T$? Probably... need not have same pooling for means as for variances, but could do as a special case instead. Want $B\Sigma_w B^T$ to have simple form, but don't see need (yet) for diagonal Σ_w. Try inverse-wishart for Σ_w combined with reduced-d that sets other wavelet scales to zero? Would have small enough number of parameters ($R \approx 15$ so Σ_w would be a (15×15) matrix).

4 Joint Modeling of Detrending and Transit Detection

The framework presented in (1) and (18) can be used to incorporate the uncertainty in detrending in a number of ways. The simplest approach, and one that can be applied to any probabilistic detrending procedure, is to replace the detrended data Y_d with different realizations from the detrending algorithm. If this is done at each iteration, it amounts to integrating over the uncertainty in the detrending. If a small number of deterministic detrenders are considered then the detrended data to be used at each iteration can be selected with equal (or weighted probability) from the list of detrended data. This approach amounts to assuming a discrete distribution for the overall mean term $\mu(t|\phi)$ in (18), thus ϕ essentially indexes the detrending routine. Note that this approach, while simple to implement, does not fully model the detrending and transit detection elements since it ignores the presence of $\mu(t|\phi)$ in the denominator of the terms in (1).

To try to model things more fully, recall that:

$$m_{sap}(t|\phi,\eta,\xi) = (\mu(t|\phi) + a(t|\xi))(1 - q(t|\eta))$$
$$m_d(t|\phi,\eta,\xi) = \left(1 + \frac{a(t|\xi)}{\mu(t|\phi)}\right)(1 - q(t|\eta)) \equiv (1 + f(t|\phi,\xi))(1 - q(t|\eta)),$$

and that:

$$f(t|\phi,\xi) = \frac{a(t|\xi)}{\mu(t|\phi)} \quad \Rightarrow \quad a(t|\xi) = \mu(t|\phi) f(t|\phi,\xi),$$

so we can rewrite (18) as:

$$Y_{sap}(t) = \mu(t|\phi)\,[1 + f(t|\phi,\xi)]\,[1 - q(t|\eta)] + \epsilon_{sap}(t|\psi), \qquad \forall\, t \in \mathcal{T}. \tag{18}$$

The algorithm described in section 4 produces posterior samples from f and q conditional on $\mu(\cdot|\phi)$, allowing us to divide out the astrophysical and transit effects, and estimate the overall instrumental trends from this 'untrended' light curve. Note that the errors in $\epsilon_{sap}(t|\psi)$ are also scaled by the divided out astrophysical signal. Fitting the original detrender to:

$$\tilde{Y}_{sap}(t) = \frac{Y_{sap}(t)}{[1 + f(t|\phi, \xi)]\,[1 - q(t|\eta)]}, \tag{19}$$

accounting for the modified error bars, will then produce an updated detrended light-curve that can be sampled from. By iterating this procedure we obtain samples from an approximation to the joint posterior distribution of detrended curves, astrophysical, transit and noise parameters. Note that this procedure ignores the dependence on ϕ in the numerator of (19), so is not completely rigorous. However, the hope is that it nevertheless captures some of the uncertainty in the detrending.