Modeling Ideas: Transit Modeling for Kepler Light Curves

Paul Baines

June 19, 2013

1 Notation & Preliminaries

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

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 $\theta = (\omega, \psi, \xi)$ and $\mathcal{T}$ be the interval on which the process is defined. We assume that:

$$\begin{aligned} Y(t|\theta) &= m(t|\omega, \xi) + \epsilon(t|\psi), \qquad \forall\, t \in \mathcal{T} \\ m(t|\omega, \xi) &= [1 - q(t|\omega)]\, f(t|\xi), \end{aligned}$$

where $\epsilon(t|\psi)$ is a mean-zero random process and $f(t|\omega)$ is the underlying intensity if there are no transits i.e., $q(t|\omega) \equiv 0$ if there are no transits. In the event of a transit we observe a version of $f(t|\xi)$ 'dampened' by a multiplicative factor $1 - q(t|\omega)$. 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:

$$\epsilon_t \sim \mathrm{GP}\left(0, \Sigma_\epsilon(\psi)\right), \tag{1}$$

where $\Sigma_\epsilon(t_i, t_j) = K(|t_i - t_j|; \psi)$ i.e., stationary covariance. Although the form in (1) is fairly rich, in practice we will often assume either independent noise (i.e., $\Sigma(t_i, t_j) = \sigma_\epsilon^2 \delta_{t_i t_j}$) or Ornstein-Uhlenbeck/AR(1)-type noise. Although we only observe a noisy version of $m(t)$, the goal is to 'separate' m 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:

$$q(t|\omega) = \left\{ \begin{array}{ll} 0 & \text{if } (t \bmod P) \notin\ [t_0, t_0 + t_d] \\ \alpha & \text{if } (t \bmod P) \in [t_0, t_0 + t_d] \end{array} \right. ,$$

where $\omega = \{t_0, t_D, \alpha, P\}$. 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, for which flexible models are required to capture all possible instrumental and astrophysical signals. We decide to model f in the wavelet domain. Let $\mathbf{w}$ denote the DWT of f and let g denote a known thresholding function (hard or soft) applied to the wavelet coeffiicients then we could model f as:

$$f(\cdot|\mathbf{w}) = \mathcal{W}^{-1}\mathbf{w}, \qquad \text{s.t.} \quad g(\mathbf{w}) = c. \tag{2}$$

For example, g could be used to constrain the high-resolution wavelet coefficients to zero. This can be just be viewed as a modeling the DWT-transformed mean as a linear function of a subset of the wavelet coefficients (i.e., or more generally, just a flexible parametric form for f parametrized by the wavelet coefficients). This full model has parameters $\theta = \{\psi, \omega, \mathbf{w}\}$, corresponding to the noise (co)variance parameters, the transit parameters and the (denoised) wavelet coefficients describing the noiseless and transit-removed signal. The full model is then just:

$$\begin{aligned} Y(t|\theta) &= \mathrm{GP}\left([1-q(t|\omega)]\, f(t|\xi), \Sigma_\epsilon(\psi)\right), \\ L(\theta|y) &= \mathcal{N}\left(y; [1-q(\cdot|\omega)]\, f(\cdot|\xi), \Sigma_\epsilon(\cdot;\psi)\right). \end{aligned} \tag{3}$$

In a likelihood framework direct maximization of (3) with respect to θ is likely to be very difficult (and given the high dimension of $\mathbf{w}$, probably not advisable). However, given there may be prior information available about certain parameters we will instead take a more flexible approach that actually leads to simpler computation. Thus, instead of applying hard constraints on the wavelet coefficients we instead propose to model:

$$\mathcal{W}\mathbf{f} = \mathbf{w}, \qquad \mathbf{w}|d, \Sigma_w \sim N\left(d, \Sigma_w\right), \qquad \mathbf{d}|\Sigma_d \sim N\left(0, \Sigma_d\right).$$

where $\mathcal{W}$ represents the DWT in matrix form. Thus:

$$\mathbf{d}|\mathbf{w}, \Sigma_w, \Sigma_d \sim N\left(\left[\Sigma_d^{-1} + \Sigma_w^{-1}\right]^{-1}\left[\Sigma_w^{-1}\mathbf{w}\right], \left[\Sigma_d^{-1} + \Sigma_w^{-1}\right]^{-1}\right).$$

This illustrates the shrinkage of the wavelet coefficients toward zero by an amount controlled by Σ_w. Typically Σ_w would be specified to ensure a large amount of shrinkage for high-resolution coefficients, and minimal shrinkage for low-resolution coefficients (and would be diagonal). In practice as well, means and variances can be pooled across resolution levels to yield:

$$w_{km}|d_k, \sigma^2_{w,k} \overset{ind}{\sim} N\left(d_k, \sigma^2_{w,k}\right), \qquad d_k \overset{ind}{\sim} N\left(0, \sigma^2_{d,k}\right), \qquad \sigma^2_{w,k} \sim \text{Inv-}\chi^2(\nu_0, s^2).$$

Thus:

$$d_k|\left\{w_{k1}, \ldots, w_{kn_k}\right\}, \sigma^2_{w,k}, \sigma^2_{d,k} \sim N\left(\frac{\frac{\sum_m w_{km}}{\sigma^2_{w,k}}}{\frac{1}{\sigma^2_{d,k}} + \frac{n_k}{\sigma^2_{w,k}}}, \frac{1}{\frac{1}{\sigma^2_{d,k}} + \frac{n_k}{\sigma^2_{w,k}}}\right).$$

For levels with little or no shrinkage we obtain:

$$d_k|\left\{w_{k1}\ldots, w_{kn_k}\right\}, \sigma^2_{w,k}, (\sigma^{-2}_{d,k} \approx 0) \overset{approx.}{\sim} N\left(\bar{w}_k, \sigma^2_{w,k}\right),$$

where $\bar{w}_k = \sum_m w_{km}$. For levels with a high degree of shrinkage, the posterior for d_k is concentrated around zero. For the time being, lets just consider Σ_d (and possibly Σ_w as well) to be specified by the analyst. Note that $\mathbf{f}$ and $\mathbf{w}$ are one-to-one so that:

This version of the full model either has parameters $\theta = \{\psi, \omega\}$ or $\theta = \{\psi, \omega, \Sigma_w\}$ depending upon how the wavelet coefficients are modeled. The likelihood is given by:

$$p(y, f, d|\theta) \propto p(y|f, \omega, \psi)p(f|d, \Sigma_w)p(d), \tag{4}$$

$$\Rightarrow \qquad p(y|\theta) = \int p(y, f, d|\theta) df\, dd, \tag{5}$$

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

$$p(f, d, \theta|y) \propto p(y|f, \omega, \psi)p(f|d, \Sigma^2_w)p(d)p(\psi, \omega, \Sigma_w), \tag{6}$$

$$\Rightarrow \qquad p(\theta|y) = \int p(f, d, \theta|y) df\, dd, \tag{7}$$

The form of (6) lends itself to a natural Gibbs sampler of the form:

$$[\psi|f, \omega], \quad [\omega|\psi, f], \quad \left[f|d, \omega, \psi, \Sigma^2_w\right], \quad \left[d|f, \Sigma^2_w\right], \quad \left[\Sigma^2_w|d, f\right],$$

where the conditioning on y throughout is dropped for brevity. Let us briefly describe each of the Gibbs updates:

- Sampling from $\psi|f,\omega$: This just amounts to sampling from the posterior distribution of the covariance parameters of a Gaussian process. If the covariance is diagonal then, with a conjugate prior, ψ can be sampled exactly (Inverse-χ^2). If ψ includes covariance parameters then this can be done using e.g., MH.
- Sampling from $\omega|\psi,f$: This will not be in closed form and other methods (e.g., MH) will be needed. Intuitively, varying ω given f varies the value of $m(t)$, so values of ω that yield m to be an appropriate mean for y (given ψ) will be favored. Likely to be the most challenging step to effectively explore the posterior.
- Sampling from $f|d,\omega,\psi,\Sigma_w^2$: Since f is sandwiched within the multilevel model, the 'prior' on f is multivariate normal, and f appears linearly in the 'likelihood', this will also be multivariate normal (albeit with some linear algebra and an inverse DWT required to compute the mean and variance).
- Sampling from $d|f,\Sigma_w^2$: This is just multivariate normal. The current value of f is transformed to the wavelet domain, and then the mean of d is given by a shrunken version of the wavelet coefficients corresponding to the shrinkage induced by the relative weights of Σ_w^2 and Σ_d^2.
- Sampling from $\Sigma_w^2|d,f$: This just amounts to sampling from the posterior distribution of the covariance parameters of a multivariate normal. If the covariance is diagonal (which it will be) or fully unstructured then, with a conjugate prior this can be sampled exactly (Inverse-χ^2). For parametrized non-diagonal matrices it can be sampled using e.g., MH.

To make things more concrete, consider the special case with n evenly-spaced osbervations, with no missing data,

$$\Sigma_\epsilon(\psi) = \sigma_\epsilon^2 I, \qquad \Sigma_d = \text{diag}\left(\sigma_{d,1}^2,\ldots,\sigma_{d,1}^2,\sigma_{d,2}^2,\ldots,\sigma_{d,J}^2\right),$$
$$\Sigma_w = \text{diag}\left(\sigma_{w,1}^2,\ldots,\sigma_{w,1}^2,\sigma_{w,2}^2,\ldots,\sigma_{w,J}^2\right),$$

where the elements of Σ_d are specified constants. Full details of the sampling procedure for this special case are given below:

- Sampling from $\sigma_\epsilon^2|f,\omega$: With a conjugate prior $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,s_0^2)$ on σ^2 we obtain the posterior:

  $$\sigma^2|f,\omega \sim \text{Inv-}\chi^2(\nu_n,s_n^2),$$

  where

  $$\nu_n = \nu_0 + n, \qquad s_n^2 = \frac{\nu_0 s_0^2 + (y-f)^T(y-f)}{\nu_n}.$$

- Sampling from $\omega|\sigma_\epsilon^2,f$: This will not be in closed form and other methods (e.g., MH) will be needed. Intuitively, varying ω given f varies the value of $m(t)$, so values of ω that yield m to be an appropriate mean for y (given ψ) will be favored.
- Sampling from $f|d,\omega,\psi,\Sigma_w^2$: Since,

  $$f|d,\Sigma_w \sim N\left(\mathcal{W}^T d, \mathcal{W}^T\Sigma_w\mathcal{W}\right), \qquad y|f \sim N\left((I-Q)f,\Sigma_\epsilon\right)$$

  Therefore:

  $$f|y,d,\Sigma_w \sim N\left(\mu_f,\Sigma_f\right),$$
  $$\Sigma_f^{-1} = \mathcal{W}^T\Sigma_w^{-1}\mathcal{W} + \frac{1}{\sigma_\epsilon^2}(I-Q)(I-Q),$$
  $$\mu_f = \Sigma_f\left[\mathcal{W}^T\Sigma_w^{-1}d + \frac{1}{\sigma_\epsilon^2}(I-Q)y\right].$$

Note that if $\Sigma_w = \sigma_w^2 I$ then Σ_f is diagonal and $\mathcal{W}^T d$ is the IDWT of d. For the more general case, $\Sigma_w^{-1} d$ is essentially a standardized vector of the mean wavelet coefficients at different scales. Equivalently, we can work with w:

$$w|d, \Sigma_w \sim N\left(d, \Sigma_w\right), \qquad y|w \sim N\left((I-Q)\mathcal{W}^T w, \Sigma_\epsilon\right).$$

Therefore:

$$w|d, \Sigma_w \sim N\left(d, \Sigma_w\right), \qquad y|w \sim N\left((I-Q)\mathcal{W}^T w, \Sigma_\epsilon\right),$$

therefore,

$$\begin{aligned} w|y, d, \Sigma_w &\sim N\left(m_w, V_f\right), \\ V_f^{-1} &= \Sigma_w^{-1} + \mathcal{W}(I-Q)\Sigma_\epsilon^{-1}(I-Q)\mathcal{W}^T, \\ m_f &= V_f\left[\Sigma_w^{-1} d + \mathcal{W}(I-Q)\Sigma_\epsilon^{-1} y\right]. \end{aligned}$$

- Sampling from $d|f, \Sigma_w^2$: Let $w = \mathcal{W}f$, then:

$$d_k|\left\{w_{k1}, \ldots, w_{kn_k}\right\}, \sigma_{w,k}^2, \sigma_{d,k}^2 \sim N\left(\frac{\frac{\sum_m w_{km}}{\sigma_{w,k}^2}}{\frac{1}{\sigma_{d,k}^2} + \frac{n_k}{\sigma_{w,k}^2}}, \frac{1}{\frac{1}{\sigma_{d,k}^2} + \frac{n_k}{\sigma_{w,k}^2}}\right).$$

  Note that both Σ_w and Σ_d are diagonal, so there are no matrix inversions in this step.

- Sampling from $\Sigma_w^2|d, f$: With independent conjugate priors on each $\sigma_{w,j}^2 \sim \text{Inv-}\chi^2(\nu_{w,j,0}, s_{w,j,0}^2)$ on each $\sigma_{w,j}^2$ for $j = 1, \ldots, J$ we obtain:

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

  where

$$\nu_{w,j,n} = \nu_{w,0} + n_j, \qquad s^2 = \frac{\nu_{w,0} s_{w,j,0}^2 + (w_j - d_j)^T (w_j - d_j)}{\nu_{w,j,n}}.$$

  Here n_j is the number of coefficients in level j of the wavelet decomposition and w_j and d_j are the subcomponents of w and d corresponding to the wavelet coefficients at level j.

2 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. The full posterior from (6) then becomes:

$$p(f, d, y_{mis}, \theta|y_{obs}) \propto p(y_{com}|f, \omega, \psi)p(f|d, \Sigma_w^2)p(d)p(\psi, \omega, \Sigma_w^2), \tag{8}$$

$$\Rightarrow \qquad p(\theta|y_{obs}) = \int p(f, d, y_{mis}, \theta|y)df\, dd\, dy_{mis}. \tag{9}$$

The form of (8) lends itself to a natural Gibbs sampler of the form:

$$[y_{mis}|y_{obs},\psi,f], \quad [\psi|f,\omega,y_{com}], \quad [\omega|\psi,f,y_{com}], \quad \left[f|d,\omega,\psi,\Sigma_w^2,y_{com}\right], \quad \left[d|f,\Sigma_w^2\right], \quad \left[\Sigma_w^2|d,f\right].$$

where the conditioning on y_{obs} and y_{mis} is now made explicit throughout. Since all previous derivations hold for the complete data, the only additional sampling step is to sample y_{mis} from the conditional posterior distribution. Fortunately, by standard normal theory, Y_{mis} is just normal. 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. 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:

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

3 Searching for Transits: Finding Candidate Period/Phase Configurations

4 Using Metadata to Improve Performance

Any ideas?

5 Model Comparison

- Ideas to 'test' for presence of transits
- Stellar variability models.
- Spacecraft event models (e.g., sudden pixel dropout)