17.1.3 Gaussian Processes Using the Spectral Approximation

This section describes a way to extend the re-expressed restricted likelihood for two-variance models to Gaussian processes (GPs). At this point, it is just an idea, though it does seem to explain two common observations about fitting GP models to data.

The idea is to use the spectral approximation to turn a GP observed on a rectangular grid into a model to which Chapter 15's tools are easily extended. The tools are applied to the approximation and not to the GP itself, but this is still of interest for two reasons. First, if we can develop interesting facts about the approximation, these facts are hypotheses about GPs, which can be tested mathematically or by simulation. Second, spectral approximations have been used to avoid the computational burden of GPs (e.g., Paciorek 2007, Fuentes & Reich 2010), so facts about the spectral approximation itself are of interest.

The presentation here draws heavily on Appendices A.1 and A.3 of Paciorek (2007), which built on Wikle (2002). My purpose is to describe the idea and give a sense of how it might help us understand GPs as data-analytic tools. Thus, I consider only the one-dimensional case, although Paciorek (2007) gives explicit expressions for two dimensions and his R package, spectralGP, handles the one- and two-dimensional cases. For the present purpose, I ignore the periodicity problem discussed in Appendix A.2 of Paciorek (2007), though a full development will need to face this problem.

For concreteness, suppose we have an updated global-mean surface temperature series with $n = 128$ observations. Paciorek (2007) requires a rectangular grid with each dimension's grid size a power of 2; for our purposes, the grid size only needs to be even. Suppose we want to smooth this series using a one-dimensional GP with no fixed effects. Some will object that this series is clearly not stationary, so it's inappropriate to fit a GP, which is stationary. I find this objection uncompelling. First, in practice GPs are routinely used to smooth data that are obviously not stationary. This may be ill-advised but standard software makes it easy and even sophisticated GP users do it. Thus, this case is of interest. Second, the material to follow is a first step toward analyses that do include fixed effects. For the global-mean surface temperature data, the fixed effects might be low-order polynomials but even if such fixed effects were included, the series's residual variation would not be stationary but merely non-stationary at higher frequencies and lower amplitude than the original series. The question then is how the residual non-stationarity affects estimates of the GP's parameters, and the following material may help us think about that.

Consider, then, a one-dimensional GP for modeling an n-vector $\mathbf{y}$, where n is even. The approximation is a mixed linear model with the intercept as the only fixed effect, so $p = 1$, with an $(n-1)$-dimensional random effect having diagonal covariance matrix $\mathbf{G}$. The equation for an observation $y_t, t = 1, \ldots, n$, is

$$y_t = \beta_0 + 2 \sum_{m=1}^{\frac{n}{2}-1} [u_{1m} \cos(\omega_m 2\pi t) - u_{2m} \sin(\omega_m 2\pi t)] + u_{1,n/2} \cos(\omega_{n/2} 2\pi t) + \varepsilon, \tag{17.25}$$

where β_0 is the intercept, the u_j are random effects, and as usual ε is an n-vector of iid $N(0,\sigma_e^2)$ errors. (This notation differs a bit from Paciorek's.) The ω_m are frequencies taking values $\omega_m \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{1}{2}\}$, indexed by $m = 1,2,\ldots,n/2$ respectively. The random-effect design matrix $\mathbf{Z}$ is $n \times (n-1)$. Its first $n-2$ columns are cos/sin pairs where the cosine and sine terms in a pair share the frequency ω_m but have different random-effect coefficients u_{1m} and u_{2m}, while $\mathbf{Z}$'s last column is an unpaired cosine term with frequency $\omega_{n/2}$. It can be shown that $\mathbf{Z}'\mathbf{Z} = n\,\mathrm{diag}(2,2,\ldots,2,1)$ and $\mathbf{1}_n'\mathbf{Z} = \mathbf{0}$. The covariance matrix $\mathbf{G}$ is diagonal, as follows. Suppose the spectral density of the GP's covariance function is $\sigma_s^2\phi(\omega;\theta)$ for frequency ω and unknown parameters θ. Then in this approximation, $u_{1,n/2}$ has variance $\frac{1}{n}\sigma_s^2\phi(\omega_{n/2};\theta)$ while u_{1m} and u_{2m} have variance $\frac{1}{2n}\sigma_s^2\phi(\omega_m;\theta)$. I consider some specific $\phi(\omega;\theta)$ below.

This approximate GP extends the two-variance model of Chapter 15 by changing the variance of each u_j from σ_s^2 times a constant to σ_s^2 times a function of the unknown θ. Because $\mathbf{Z}$'s columns are orthogonal and $\mathbf{X}'\mathbf{Z} = \mathbf{0}$, the desired simple form of the restricted likelihood is easily derived. The approximate model is

$$\mathbf{y} = \mathbf{1}_n\beta + \mathbf{Z}\mathbf{u} + \varepsilon \quad \text{for } \varepsilon \sim N(\mathbf{0},\sigma_e^2\mathbf{I}_n) \text{ and } \mathbf{u} \sim N(\mathbf{0},\sigma_s^2\mathbf{D}(\theta)), \tag{17.26}$$

where $\mathbf{D}(\theta)$ is $(n-1)\times(n-1)$ and diagonal, as above. Pre-multiply (17.26) by $(\mathbf{Z}'\mathbf{Z})^{-0.5}\mathbf{Z}'$ — note that the power is –0.5, not –1 — to give

$$\begin{aligned} \hat{\mathbf{v}} \equiv (\mathbf{Z}'\mathbf{Z})^{-0.5}\mathbf{Z}'\mathbf{y} &= (\mathbf{Z}'\mathbf{Z})^{0.5}\mathbf{u} + (\mathbf{Z}'\mathbf{Z})^{-0.5}\mathbf{Z}'\varepsilon \\ &\equiv \mathbf{v} + \xi, \end{aligned} \tag{17.27}$$

where ξ is $(n-1)\times 1$ with $\mathrm{cov}(\xi) = \sigma_e^2\mathbf{I}_{n-1}$ and $\mathbf{v}$ is $(n-1)\times 1$ with

$$\begin{aligned} \mathrm{cov}(\mathbf{v}) &= \sigma_s^2(\mathbf{Z}'\mathbf{Z})^{0.5}\mathbf{D}(\theta)(\mathbf{Z}'\mathbf{Z})^{0.5} \\ &= \sigma_s^2\mathrm{diag}(\ \phi(\omega_{m(j)};\theta)\ ) \\ \text{for } m(j) &= \left\{ \begin{array}{l} (j+1)/2 \text{ for odd } j \\ j/2 \text{ for even } j \end{array} \right\}, j = 1,\ldots,n-1. \end{aligned}$$

The restricted likelihood is the likelihood for $(\sigma_e^2,\sigma_s^2,\theta)$ arising from (17.27):

$$\begin{aligned} \log RL(\sigma_e^2,\sigma_s^2,\theta) &= K - 0.5\sum_{j=1}^{n-1}\left[\ \log(\sigma_s^2 a_j(\theta)+\sigma_e^2) + \hat{v}_j^2/(\sigma_s^2 a_j(\theta)+\sigma_e^2)\ \right] \\ \text{where } a_j(\theta) &= \phi(\omega_{m(j)};\theta), j = 1,\ldots,n-1, \end{aligned} \tag{17.28}$$

and K is an unimportant constant.

The restricted likelihood for the approximate model has no free terms for σ_e^2. The mixed terms are

$$-0.5\left[\ \log(\sigma_s^2 a_j(\theta)+\sigma_e^2) + \hat{v}_j^2/(\sigma_s^2 a_j(\theta)+\sigma_e^2)\ \right]; \tag{17.29}$$

$j = 1,2$ correspond to frequency ω_1, $j = 3,4$ to frequency ω_2, and so on. The canonical predictors are the columns of $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-0.5}$ and the $\hat{v}_j$ are the coefficients of projections of $\mathbf{y}$ onto those columns. Thus the $\hat{v}_j$ decompose $\mathbf{y}$ into components corresponding to the frequencies ω_m. Note that the canonical predictors are the same for all GPs on a given rectangular grid, so for a given $\mathbf{y}$ the $\hat{v}_j$ are also the same for all GPs. Therefore, given $\mathbf{y}$, the restricted likelihoods for different GPs are differentiated — in this approximation — *only* by their $a_j(\theta)$.

Because $a_j(\theta)$ is a function of the unknown θ, this scalarized restricted likelihood is not a gamma-errors generalized linear model with identity link. However, given θ, this spectral approximation is a two-variance model and its restricted likelihood is a gamma-errors GLM as in Chapters 15 and 16. In those chapters, the a_j were the key to understanding the restricted likelihood, so an understanding of how a GP works as a data-analysis engine (using the spectral approximation) begins with examining the $a_j(\theta)$ as a function of θ.

To test this idea's potential, we'll consider two GPs in the Matérn family. Paciorek (2007, equation 5) parameterizes the Matérn family for D-dimensional space somewhat differently than usual, so that the spectral density is

$$\sigma_s^2\phi(\omega;\rho,\nu) = \sigma_s^2 K(\nu,D)\rho^D\left(1+\frac{(\pi\rho)^2}{4\nu}\omega'\omega\right)^{-(\nu+D/2)}, \qquad (17.30)$$

where ν and ρ are the Matérn family's smoothness and range parameters, ω is the D-vector of frequencies, and $K(\nu,D)$ is a function of ν and D but not ρ or ω. This $\phi(\omega;\rho,\nu)$ has the form of a D-variate t density with dispersion matrix proportional to the identity and scale parameter $\sqrt{2}/\pi\rho$. For the one-dimensional case, $D = 1$ and $\omega'\omega = \omega^2$. In the Matérn model, ρ describes the range at which the correlation between pairs of observations (t_1,t_2) decays to a small value.

The parameter ν is usually described as controlling the smoothness of realizations generated from the GP when it is used as a probability model (as distinct from using it as a likelihood) and ν is usually fixed *a priori* in analyses. We'll consider two GPs, with $\nu = 0.5$, the exponential form, and $\nu = \infty$, the squared exponential form. For $\nu = 0.5$, $\phi(\omega;\rho)$ has the form of a Cauchy density,

$$\phi(\omega;\rho) = \frac{1}{\sqrt{2}}\rho\left(1+\frac{(\pi\rho)^2}{2}\omega^2\right)^{-1}, \qquad (17.31)$$

while for $\nu = \infty$, $\phi(\omega;\rho)$ has the form of a normal (Gaussian) density,

$$\phi(\omega;\rho) = \frac{\sqrt{\pi}}{2}\rho\exp\left(-\frac{(\pi\rho)^2}{4}\omega^2\right). \qquad (17.32)$$

The functions $\phi(\omega;\rho)$ give the $a_j(\rho)$ in the re-expressed restricted likelihood for these two (approximate) GP models. How do these $a_j(\rho)$ decline as the frequency ω_m increases, and how does this differ between $\nu = 0.5$ and $\nu = \infty$? Figure 17.1 tells the story. Figure 17.1 shows $a_j(\rho)$ for two values of ρ, 0.1×128 (i.e., 10% of the series length) and 0.025×128. However, the pattern is the same for any ρ. For a given ρ, the two models ($\nu = 0.5$ and $\nu = \infty$) have quite similar $a_j(\rho)$, and these $a_j(\rho)$ are the *only* aspect of the restricted likelihood that differs between these two (approximate) GP models. This explains the common observation that in fitting a GP to data, the data provide hardly any information about ν. The two ν depicted here

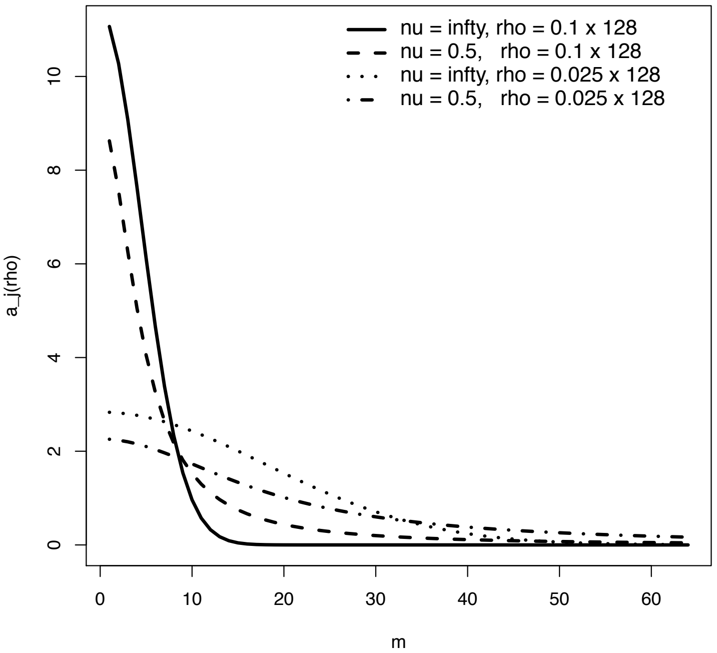

Figure 17.1: For a global-mean surface temperature series with $n = 128$, $a_j(\rho)$ for GPs with $\nu = 0.5$ and $\nu = \infty$. The horizontal axis is m, the index of the frequencies ω_m. (This choice of axis avoids showing repeated $a_j(\rho)$.)

are the conventional extreme small and large values of ν and they produce restricted likelihoods that barely differ, so it is no wonder that data cannot distinguish more subtle differences in ν.

Figure 17.1 shows that the $a_j(\rho)$ do depend strongly on ρ. Recalling that for fixed ρ this model is a gamma-errors GLM with identity link but that ρ can also be adjusted, we can develop some intuition for how the unknown ρ, σ_s^2, and σ_e^2 are adjusted to fit the data. In this gamma-errors GLM, the "observations" are the $\hat{v}_j^2$, which have expectation

$$E(\hat{v}_j^2|\sigma_s^2, \rho, \sigma_e^2) = \sigma_s^2 a_j(\rho) + \sigma_e^2. \tag{17.33}$$

For large j, $E(\hat{v}_j^2|\sigma_s^2, \sigma_e^2) \approx \sigma_e^2$, so select a value for $\hat{\sigma}_e^2$ in the "middle" of the $\hat{v}_j^2$ for large j. Recall that $a_j(\rho) = \rho f(\rho, \omega_{m(j)}^2)$, where f declines as j increases; so choose $\hat{\rho}$ to fit the rate at which the $\hat{v}_j^2$ decline for small j. Finally, choose $\hat{\sigma}_s^2$ to make $\hat{\sigma}_s^2 a_j(\hat{\rho}) + \hat{\sigma}_e^2$ go through the middle of the $\hat{v}_j^2$ for small j.

It is now easy to understand the common observation that σ_s^2 and ρ are poorly identified. The only thing identifying them is the rate at which $f(\rho, \omega_m^2)$ declines and if we recall the "noise" in $\hat{v}_j^2$ as a function of j for earlier problems we've considered, it is clear that ρ is not very well-determined by the data, so σ_s^2 isn't, either.

To suggest how this approximation might be used to learn about GPs, I offer some conjectures about how influential $\hat{v}_j^2$ — squared lengths of projections onto particular sine or cosine basis functions — might affect estimates of ρ, σ_s^2, and σ_e^2. These conjectures can be tested for actual GPs using simulation experiments. (An exercise incites you to do so.)

One conjecture begins with the observation that a strong low-frequency trend (e.g., quadratic) implies large $\hat{v}_j^2$ for small j but much smaller $\hat{v}_j^2$ for succeeding j. To capture this sharp decline in the $\hat{v}_j^2$, $\hat{\rho}$ will be large. Then $\hat{\sigma}_s^2$ will be selected so that for small j, $\hat{\sigma}_s^2 a_j(\hat{\rho}) + \sigma_e^2$, which is dominated by $\hat{\sigma}_s^2 a_j(\hat{\rho})$, fits the large $\hat{v}_j^2$. The resulting fit to $\mathbf{y}$ will, however, not necessarily be smooth. That depends on the level of the $\hat{v}_j^2$ for large j, which determines $\hat{\sigma}_e^2$ and thus the DF in the fit of a canonical predictor's coefficient v_j, which is

$$\frac{a_j(\rho)}{a_j(\rho) + \sigma_e^2/\sigma_s^2} = \frac{\sigma_s^2 a_j(\rho)}{\sigma_s^2 a_j(\rho) + \sigma_e^2}. \tag{17.34}$$

If $\hat{\sigma}_e^2$ is large relative to $\hat{\sigma}_s^2 a_j(\hat{\rho})$, the fit is smooth; otherwise, the fit is rough.

Another conjecture is that an outlier in a high frequency will have little effect on $\hat{\sigma}_s^2$ or $\hat{\rho}$ but will inflate $\hat{\sigma}_e^2$ and thus result in a smoother fit.

17.1.4 Separable Models

The restricted likelihood has the simple re-expression for models that are separable in the following sense. Suppose the n-vector $\mathbf{y}$ is modeled as $\mathbf{y} = \delta + \varepsilon$, where $\varepsilon \sim N_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ and δ has a possibly improper normal density with mean zero and precision matrix of the form $\sum_{k=1}^K \tau_k \mathbf{Q}_k$, where the τ_k are scalar precision parameters and the $\mathbf{Q}_k$ are $n \times n$ matrices. The $\mathbf{Q}_k$ have the form $\mathbf{Q}_k = \mathbf{A}_1 \otimes \ldots \otimes \mathbf{A}_M$, where $\otimes$ is the Kronecker product, $\mathbf{A}_l = \mathbf{I}$ for $l \neq k$, $\mathbf{A}_k$ is positive semi-definite, and $\mathbf{A}_l$ has the same dimensions for all $\mathbf{Q}_k$. The proof is straightforward and given as an exercise; the idea is that by properties of the Kronecker product, we just need to simultaneously diagonalize the identity and one positive semi-definite $\mathbf{A}_k$ at a time, and this can always be done.

This class of models includes some 2NRCAR models, for example, those in which the areas form a rectangular grid with neighbors within rows being one class of neighbors and neighbors within columns being the second class (Besag & Higdon 1999). This class of models also includes spatio-temporal ICAR models in which spatial neighbors are one class of neighbors and temporal neighbors are the second class. Spatio-temporal ICAR models with more than one class of spatial neighbor pairs are separable if the spatial part of the model is.

The following two assertions appear to be true in general for separable models as defined above but I have not proved them (proofs are exercises). This form of separability is a type of balance but separable models as defined here do not satisfy the conditions of general balance. Also, models of this form do not produce restricted likelihoods that are likelihoods for generalized linear models.

The intuition for these assertions comes from a separable 2NRCAR model. Re-