

Efficient Emulators of Computer Experiments Using Covariance Tapering

Cari Kaufman and Derek Bingham

February 26, 2007

1 Emulators Using Gaussian Process Models

- $Y(x)$, $x \in \mathbb{R}^p$ output of computer code
- observations $Y(x_1), \dots, Y(x_n)$
- Model $Y(x) = \beta^T f(x) + Z(x)$, Z having Gaussian process distribution with mean zero and covariance function

$$\text{Cov}(Z(x), Z(x')) = \sigma^2 R(x, x'; \theta)$$

- In typical implementations, the correlation has the form

$$R(x, x'; \theta) = \prod_{j=1}^d R_j(|x_j - x'_j|; \theta_j), \quad (1)$$

where $R_j(\cdot; \theta_j)$ is some smoothly decreasing, strictly positive covariance function, such as $R_j(w) = \exp\{-\theta w^p\}$.

- See e.g. Sacks et al. (1989), Welch et al. (1992), and Kennedy and O'Hagan (2001)
- We want to
 - Predict $Y(x^*)$ at unobserved input x^*
 - Estimate the prediction error variance
 - (Estimate β, σ^2, θ)
- We can achieve these by
 - Estimating β, σ^2, θ by maximum likelihood and plugging into the BLUP

$$\hat{Z}(x^*) = \hat{\beta}^T f(x^*) + \gamma(\hat{\theta})^T \Gamma(\hat{\theta})^{-1} (Y - X\hat{\beta}) \quad (2)$$

where $\Gamma(\hat{\theta})$ is the estimated correlation matrix of the observations Y and $\gamma(\hat{\theta})$ is the estimated correlation between Y and $Y(x^*)$

- Deriving a joint posterior distribution for β, σ^2, θ , and $Y(x^*)$
- Estimating the parameters by maximum likelihood, then fixing them in the Bayesian analysis.

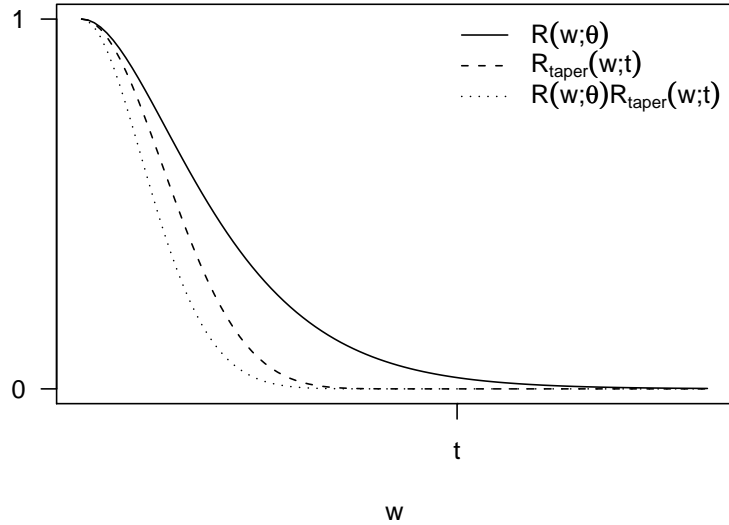
- In any case, we are dealing with expressions as in (2), as well as the likelihood

$$\mathcal{L}(\beta, \sigma^2, \theta) = (2\pi\sigma^2)^{-n/2} |\Gamma(\theta)|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (Y - X\beta)^T \Gamma(\theta)^{-1} (Y - X\beta) \right\}$$

- Problem: The determinant and inverse both require $O(n^3)$ operations, so are infeasible for n greater than a few thousand.

2 Covariance Tapering

- Covariance tapering: Replace $R(\cdot; \theta)$ by $R(\cdot; \theta)R_{\text{taper}}(\cdot; t)$, where $R_{\text{taper}}(\cdot; t)$ is a positive definite function which is identically zero for values greater than t .



- The resulting matrix $\Gamma(\theta) \circ T(t)$ is sparse and can be manipulated more efficiently.
- Working in the framework of spatial datasets, Furrer et al. (2006) showed that one can obtain asymptotically optimal interpolations using tapering, replacing $\gamma(\theta)^T \Gamma(\theta)^{-1} (Y - X\beta)$ by $[\gamma(\theta) \circ T^*(t)]^T [\Gamma(\theta) \circ T(t)]^{-1} (Y - X\beta)$. Note that this assumes the model parameters are known.
- Kaufman (2006) showed that certain covariance parameters can still be consistently estimated by maximizing one of two approximations to the likelihood using tapering.

(This assumes the mean is known to be zero.)

$$\begin{aligned}\ell_{t1aper}(\sigma^2, \theta) &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |\Gamma(\theta) \circ T(t)| \\ &\quad - \frac{1}{2\sigma^2} Y^T [\Gamma(\theta) \circ T(t)]^{-1} Y \\ \ell_{t2apers}(\sigma^2, \theta) &= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |\Gamma(\theta) \circ T(t)| \\ &\quad - \frac{1}{2\sigma^2} Y^T \left([\Gamma(\theta) \circ T(t)]^{-1} \circ T(t) \right) Y\end{aligned}$$

Maximizing $\ell_{t2apers}$ corresponds to solving an unbiased estimating equation and in practice gives estimators with much smaller mean squared error. However, for purposes of prediction, it may work better to use ℓ_{t1aper} , since the tapered covariance will be used in the BLUP.

- Both Furrer et al. (2006) and Kaufman (2006) considered isotropic covariance functions. In this case, the form of the covariance can be chosen to satisfy conditions for these theoretical results, while the range of the covariance, t , should be chosen as large as is computationally feasible, as larger values of t correspond to smaller variability in the estimators/predictors.
- When the covariance has the product form (1), the question of how to choose the taper function T is more complicated. For instance, intuition suggests we want might to taper less severely in the dimensions with high correlation.
- We consider tapering individually in each dimension:

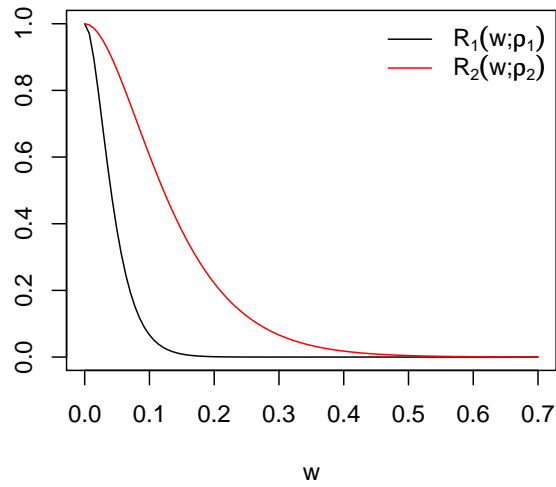
$$R(x, x'; \theta) R_{taper}(x, x'; t) = \prod_{j=1}^d R_j(|x_j - x'_j|; \theta_j) R_{taper,j}(|x_j - x'_j|; t_j)$$

- A possible strategy: First use a subset of the total dataset to get preliminary estimates of the correlation functions R_j . Then adapt the taper range t_j to the range of the estimated covariance in each coordinate.

3 Simulation Study

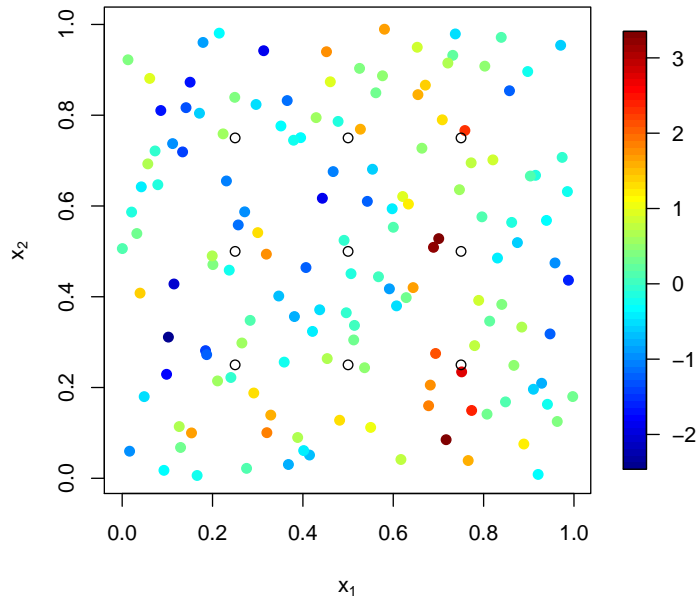
- Carry out a simulation study with 2D input and different correlation lengths: is it worth “tuning” t_1 and t_2 ?
- Simulate 100 datasets with
 - locations from a latin hypercube design on the unit square, $n = 150$
 - $\beta = 0$; $R_j(w; \rho_j)$ Matérn with smoothness parameter $\nu = 2$:

$$R_j(w; \rho_j) = \frac{(w/\rho_j)^\nu}{\Gamma(\nu)2^{\nu-1}} \mathcal{K}_\nu(w/\rho_j)$$



– $\rho_1 = 0.02, \rho_2 = 0.06$

- Example dataset:



Estimate parameters based on $n = 150$ observations (colored dots); predict at nine sets of input values (empty dots).

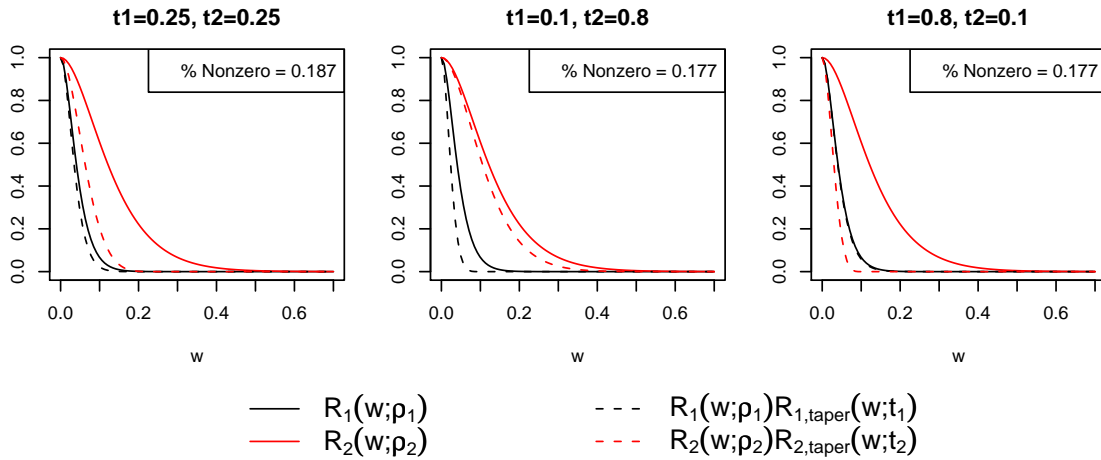
- For each dataset, maximize likelihood and the tapering approximation

$$\begin{aligned} \ell_{1taper}(\sigma^2, \rho_1, \rho_2) = & -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |\Gamma(\rho_1, \rho_2) \circ T(t_1, t_2)| \\ & - \frac{1}{2\sigma^2} Y^T [\Gamma(\rho_1, \rho_2) \circ T(t_1, t_2)]^{-1} Y \end{aligned}$$

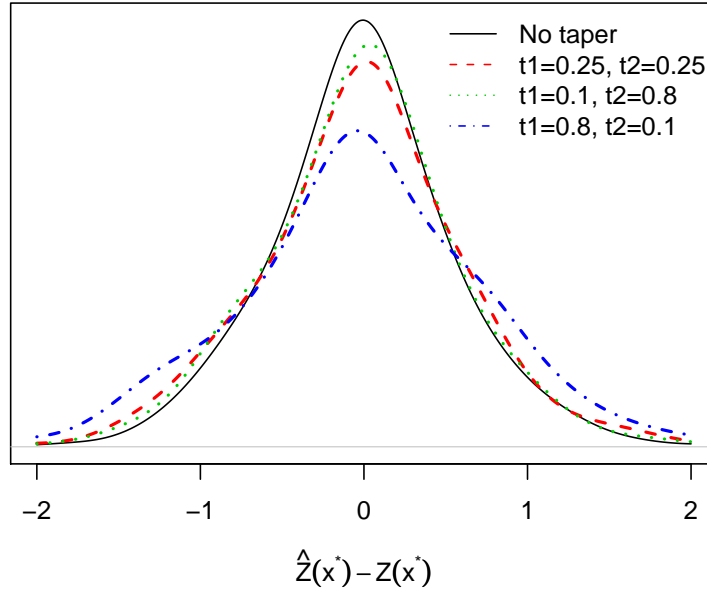
- Then predict the output at the nine locations above using the BLUP with the un-tapered or tapered covariance matrix, plugging in the corresponding parameter estimates.
- Use a Wendland tapering function in each coordinate

$$R_{i,taper}(w; t_i) = \left(1 - \frac{w}{t_i}\right)^8 \left(32 \left(\frac{w}{t_i}\right)^3 + 25 \left(\frac{w}{t_i}\right)^2 + 8 \left(\frac{w}{t_i}\right) + 1\right) I(w < t_i)$$

- The simplest choice is to choose a degree of sparsity which makes the computations feasible, then find a single value t such that the taper matrix $T(t_1 = t, t_2 = t)$ has that degree of sparsity.
- However, by increasing one taper range and decreasing the other, we can achieve the same degree of sparsity.
- Choices of tapering functions: single taper range, “properly tuned,” and “poorly tuned”



- Kernel density estimates of prediction errors:



- Some theoretical questions

- Can we extend the theoretical results for isotropic correlations to this case? For example, a condition for equivalence of two mean zero Gaussian measures with spectral densities f_1 and f_2 is that there exists some c such that

$$\int_{\|\omega\|>c} \left\{ \frac{f_1(\omega) - f_0(\omega)}{f_0(\omega)} \right\} d\omega < \infty,$$

(Stein, 2004). If f_1 and f_2 are isotropic, we can rewrite the integral in polar coordinates, and it suffices to show

$$\left| \frac{f_1(ru)}{f_0(ru)} - 1 \right| = O(r^{-\xi}) \text{ for some } \xi > d/2.$$

- What tapering functions are appropriate for correlations of the form $R(w) = \exp\{-\theta w^p\}$? In particular, $p = 2$ gives the so-called Gaussian covariance, which is also the limiting case of the Matérn covariance as $\nu \rightarrow \infty$. However, existing results indicate that the degree of the Wendland polynomial must increase with ν (Kaufman, 2006).
- Some thoughts on future directions
 - Comparing the estimates of mean squared prediction error. For example, what are the coverage probabilities of confidence intervals for the predicted values?

- Testing on a known function, rather than realizations of a stochastic process. (What are the correlation lengths in a typical example, relative to the domain of the observations?)
- Forming preliminary estimates based on a subsample
- Is it worth using the tapered Matérn as the model, or once we have estimates of the correlation length, can we substitute the taper functions themselves as the correlations?
- Application to astrophysics example

References

- Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15:502–523.
- Kaufman, C. G. (2006). *Covariance Tapering for Likelihood Based Estimation in Large Spatial Datasets*. PhD thesis, Carnegie Mellon University.
- Kennedy, M. and O’Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B (Methodology)*, 64:425–464.
- Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4:409–435.
- Stein, M. (2004). Equivalence of Gaussian measures for some nonstationary random fields. *Journal of Statistical Planning and Inference*, 123:1–11.
- Welch, W., Buck, R., Sacks, J., Wynn, H., Mitchell, T., and Morris, M. (1992). Screening, prediction, and computer experiments. *Technometrics*, 34:15–25.