Efficient Emulators of Computer Experiments Using Covariance Tapering

Cari Kaufman

January 29, 2007

1 Emulators Using Gaussian Process Models

- $Y(x), x \in \Re^p$ output of computer code
- observations $Y(x_1), \ldots, Y(x_n)$
- As in Welch et al. (1992), model $Y(x) = \beta + Z(x)$, Z having Gaussian process distribution with mean zero and covariance function

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

- 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

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

    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 (1), as well as the likelihood

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

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

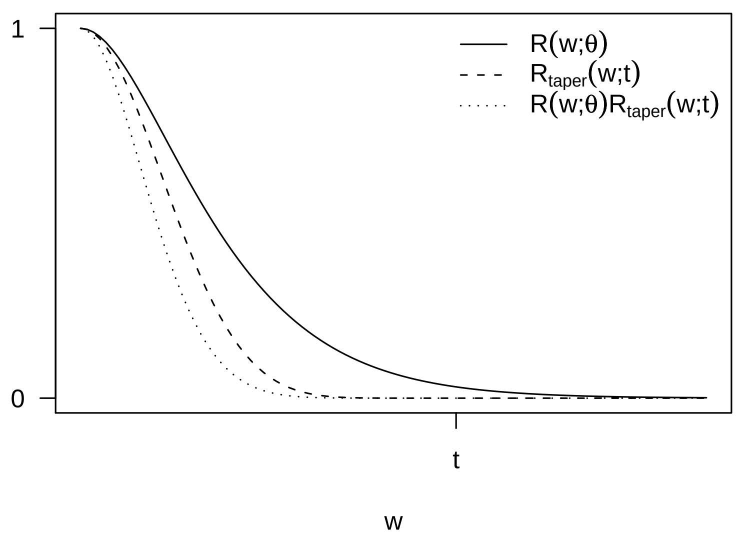
2 Covariance Tapering

- In typical implementations, the correlation typically has the form

$$R(x, x') = \prod_{j=1}^{d} R_j(|x_j - x'_j|; \theta_j), \tag{2}$$

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

- Covariance tapering: Replace $R_j(\cdot; \theta_j)$ by $R_j(\cdot; \theta_j) R_{taper,j}(\cdot; t_j)$, where $R_{taper,j}(\cdot; t_j)$ is a positive definite function which is identically zero for values greater than t_j.



- 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, and Kaufman (2006) showed that certain covariance parameters can still be consistently estimated by maximizing an approximation to the likelihood using tapering. However, both of these considered isotropic covariance functions.

- When the covariance has the form (2), the question of how to choose the taper function T_j is more complicated. For instance, intuition suggests we want might to taper the "important" input dimensions less severely.

- 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

- To use this strategy, we first need to develop "rules of thumb" for cases in which the true covariance functions are known.
- Carry out two simulation experiments with 2D input: one in which the covariance ranges are the same but the taper ranges are different, and one in which the covariance ranges are different but the taper ranges are the same.
- For each experiment, simulate 100 datasets with
  - locations a 10 by 10 grid over the unit square
  - $\beta = 0$; $R_j(w; \rho_j) = \exp\{-w/\rho_j\}$
- For each dataset, maximize likelihood and the tapering approximation

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

- Taper function is a product of Wendland functions

$$R_{taper}(x, x') = R_{1,taper}(|x_1 - x_1'|; t_1) R_{2,taper}(|x_2 - x_2'|; t_2),$$

  with

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

- Two experiments:
  1. Same range, different taper: $\rho_1 = 0.4, \rho_2 = 0.4, t_1 = 10, t_2 = 0.2$
  2. Different range, same taper: $\rho_1 = 0.2, \rho_2 = 0.6, t_1 = 0.4, t_2 = 0.4$
- Our expectations:
  - Relatively small bias, although perhaps not due to change in isotropy. (Estimators are asymptotically unbiased.)
  - Coordinate which is tapered more relative to its correlation range will have greater variance in estimator of ρ.

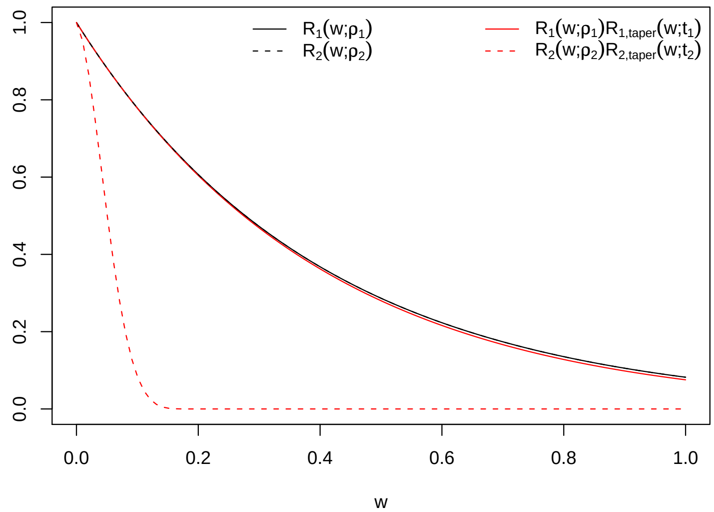


Figure 1: Original and tapered covariance functions, when $\rho_1 = 0.4, \rho_2 = 0.4, t_1 = 10$, and $t_2 = 0.2$.

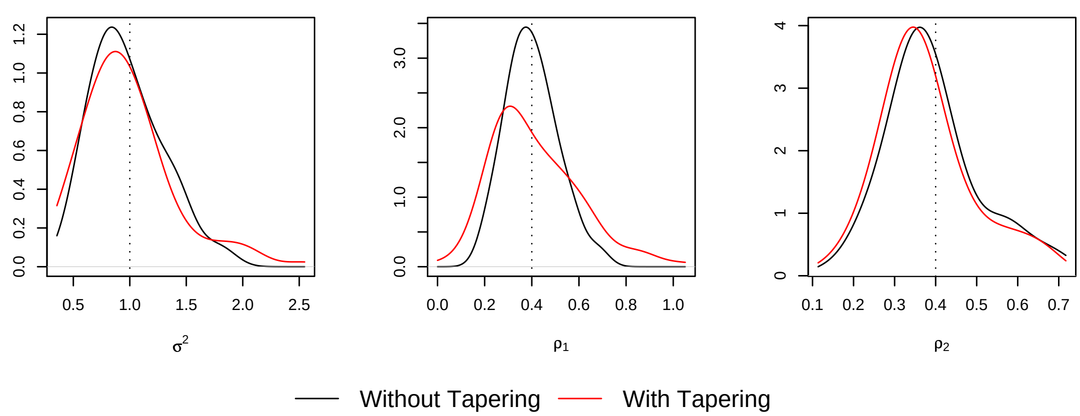


Figure 2: Kernel density estimates of sampling densities of estimators, when $\rho_1 = 0.4, \rho_2 = 0.4, t_1 = 10$, and $t_2 = 0.2$.

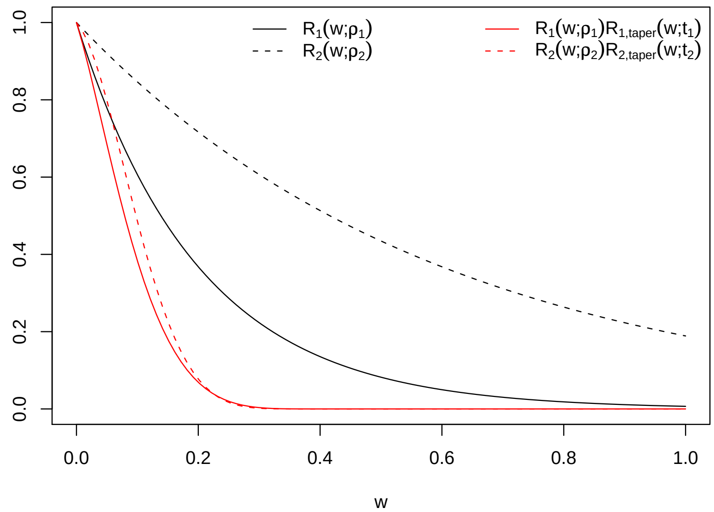


Figure 3: Original and tapered covariance functions, when $\rho_1 = 0.2, \rho_2 = 0.6, t_1 = 0.4$, and $t_2 = 0.4$.

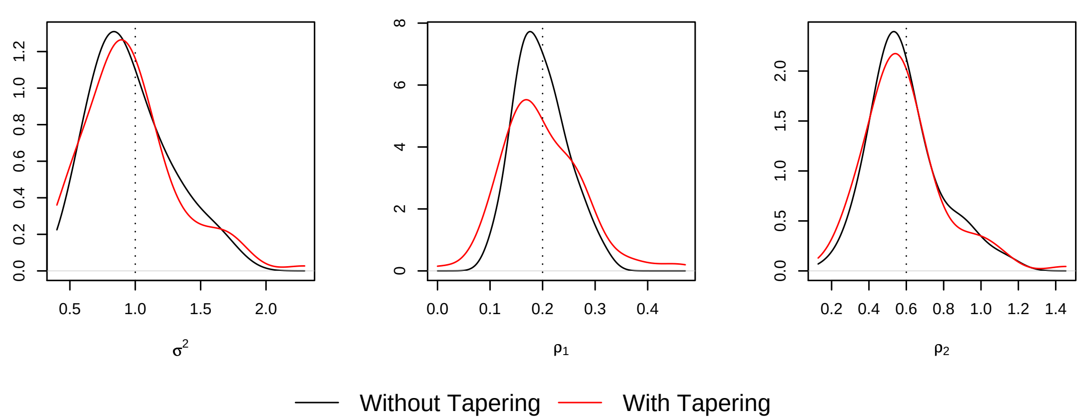


Figure 4: Kernel density estimates of sampling densities of estimators, when $\rho_1 = 0.2, \rho_2 = 0.6, t_1 = 0.4$, and $t_2 = 0.4$.

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.

Welch, W., Buck, R., Sacks, J., Wynn, H., Mitchell, T., and Morris, M. (1992). Screening, prediction, and computer experiments. *Technometrics*, 34:15–25.