

Facile Spacio-Temporal Modeling for Massive Data: WIDALS, Weighting by Inverse Distance with Adaptive Least Squares

Dave Zes
Department of Statistics
UCLA

2011-11-12

ABSTRACT. In the following we unite Adaptive Least Squares (ALS) and Inverse Distance Weighting as a computationally frugal means of modeling very large space-time data. This technique, dubbed weighting by inverse distance with adaptive least squares (WIDALS) boasts several merits, including a small and readily interpretable hyperparameter space, and relative ease of implementation. An example of interpolation of global daily maximum temperatures is included.

1 Introduction

The sub-discipline of statistics that focuses on observations over the space-time domain has received considerable interest recently. Analogous *systems* have undergone a parallel development in engineering, where much emphasis is given to guidance and signal tracking. While differences in application between the two disciplines necessitates unique attention to context, fruitful discussions of methodology in either field necessitates appreciating analogous methods known to the other. It is also true that particular analytical and computational challenges that arise in one will also at some point arise in the other, since both ultimately call upon the same theoretical underpinnings. The growing availability of massive data spawns one such challenge. While one on hand the statistician performing inference is attracted by large n as a means of increasing the precision of sought point estimates or predictions, there lurks the inhibition that the computational expense of an exact linear (or a translated non-linear) solution is $O(n^3)$ [5] (p 673). Moreover, exactly solving a solution for large n potentially introduces numerical instability. In engineering, attention has come in the form of variants of certain solutions. The square-root Kalman filter, for example, mitigates numeric instability [3, 9], and its “unscented” variant reduces computational complexity of retrospective hyperparameterization, but the system solution is still $O(n^3)$ [12]. In the statistical setting, for example when examining terrestrial processes, there commonly exists the constraint that observations over space covary as some well behaved function of distance (a “covariogram”).

This constraint, in-and-of-itself, however, does not mitigate the computational burden — although, ordered spacial location of sensors can mitigate the computational burden by using iterative matrix inversion techniques, e.g., [1].

In our statistical setting, much emphasis, therefore, has been on approximating an ideal solution. Likely the most popular *spacial* prediction technique, at least historically, has been to interpolate (predict to an unmonitored site) using a linear combination of recorded observations and weights created directly as some function of distance, a technique generically termed “inverse distance weighting.” Another common approach interpolates using a local neighborhood of observations. This $O(n)$ approach has some rigorous justification. Michael Stein, a prolific contributor to the field, shows that, given certain assumptions about the covariance function and the location of sites, under asymptotic infill the normal system weights are dominated by nearby locations — the ratio of the MSE over the ideal MSE converges to unity [10]. The effect of distant observations are, in essence, “screened” by local observations — a conjecture that had been erstwhile empirically supported in geostatistics [4]. The discontinuity of spacial predictions typically produced by this method should not be considered a shortcoming; however, if we seek to interpolate to n^* unmonitored locations, this will mean the sequential inversion of all the small concomitant local observation matrices — perhaps n or so inversions if $n^* > n$. Another promising means of approximating a linear solution is through sparse matrix representation [8]. In the spacio-temporal context, Furrer, Genton, and Nychka have advanced the idea of “tapering” [2], zeroing-out small estimated covariation in the responses across the spacial domain (which could certainly also be extended to more adventurous efforts to model covariograms over both the spacial and temporal domains).

Weighting by Inverse Distance with Adaptive Least Squares (WIDALS) proposed in the present work offers none of the elegance or ingenuity of the afore-mentioned contributions. WIDALS is a brute force union between Adaptive Least Squares (ALS) and inverse distance weighting — both, by themselves, well known to the literature. For the sake of brevity we apply WIDALS to a single data set. However, over the past couple years I have applied WIDALS to dozens of real datasets and hundreds of simulations, where it has exhibited numerous virtues.

- ❶ Computational frugality (e.g., forecasting or interpolating to fixed n^* sites at τ time steps is $O(n\tau)$).
- ❷ Implicit handling of temporal correlation of model noise.
- ❸ Small hyperparameter space (4-5 readily interpretable non-negative scalar values).
- ❹ Hyperparameter robustness. Miss-location of hyperparameters in fitting typically results in relatively mild deterioration of testing RMSE.

⑤ Because of the small hyperparameter space, stochastic fitting techniques, such as Metropolis Hastings (MH) can be employed.

⑥ MH allows for minimization of an arbitrary cost function, and is perfectly suited for parallel processing.

⑦ Because the WIDALS solution explicitly segregates the deterministic and stochastic parts, it is readily extensible, e.g., the researcher can easily adapt the algorithm to account for time-varying spacial covariation.

Symbol	Definition
ω	row vector of n spacial locations
\mathbf{Z}	$\tau \times n$ response data matrix
\mathbf{z}_t	row vector of n responses at time t
\mathbf{H}_t	$n \times p$ matrix of covariates
β_t	column vector of p partial slopes that map \mathbf{H}_t to \mathbf{z}_t
\mathbf{b}_t	an estimator for β_t
\mathbf{D}^Ω	the hollow $n \times n$ matrix of spacial distances between ω

Table 1: Notational Symbols

2 ALS to Locate an Explicit System Mean Function

For illustration, suppose our observations arise through

$$\beta_t = \beta_{t-1} + \nu_t \tag{1}$$

$$\mathbf{z}_t^T = \mathbf{H}_t \beta_t + \varepsilon_t \tag{2}$$

with $\nu_t \sim \mathcal{N}[\mathbf{0}, \mathbf{Q}_t]$ and $\varepsilon_t \sim \mathcal{N}[\mathbf{0}, \mathbf{R}_t]$. This system may readily be cast as a time-varying parameter regression problem, where β_t is our $p \times 1$ parameter (i.e., partial slopes), and \mathbf{H}_t , ($n \times p$), is our known regressor (or *design*) matrix at time t .

2.1 1-Step Ahead Forecast

Suppose the general case where the observation errors are cross-correlated across time, e.g., $\mathbf{R}_t^{(1)} = \text{Cov}[\varepsilon_t, \varepsilon_{t-1}]$, and $\mathbf{R}_t^{(t-1)} = \text{Cov}[\varepsilon_t, \varepsilon_1]$, etc., and we seek $\mathbf{E}[\mathbf{z}_t | \mathbf{Z}_{1:t-1}]$.

We first define

$$\boldsymbol{\Sigma}^\times := \begin{pmatrix} \mathbf{R}_t^{(0)} & \mathbf{R}_1^{(t)} & \dots & \mathbf{R}_t^{(t-2)} \\ \mathbf{R}_t^{(1)} & \mathbf{R}_t^{(0)} & \dots & \mathbf{R}_t^{(t-3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_t^{(t-2)} & \mathbf{R}_t^{(t-3)} & \dots & \mathbf{R}_t^{(0)} \end{pmatrix}, \quad \boldsymbol{\Sigma}^\star := \begin{pmatrix} \mathbf{R}_t^{(t-1)} \\ \mathbf{R}_t^{(t-2)} \\ \vdots \\ \mathbf{R}_t^{(1)} \end{pmatrix} \quad (3)$$

and

$$\boldsymbol{\zeta}_{1:t-1} := (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{t-1}) \quad (4)$$

If we create a vertically stacked regressor matrix

$$\boldsymbol{\Gamma}_{t-1} := \begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_{t-1} \end{pmatrix} \quad (5)$$

The LS BLUP comes by way of

$$\mathbf{b} = (\boldsymbol{\Gamma}_{t-1}^T \boldsymbol{\Sigma}^\times \boldsymbol{\Gamma}_{t-1})^{-1} \boldsymbol{\Gamma}_{t-1}^T \boldsymbol{\Sigma}^\times \boldsymbol{\zeta}_{1:t-1} \quad (6)$$

$$\widehat{\boldsymbol{\zeta}}_{1:t-1}^T = \boldsymbol{\Gamma}_{t-1}^T \mathbf{b} \quad (7)$$

$$\widehat{\mathbf{z}}_t = \mathbf{b}^T \mathbf{H}_t^T \quad (8)$$

$$\widetilde{\mathbf{z}}_t = \widehat{\mathbf{z}}_t + \left(\boldsymbol{\zeta}_{1:t-1} - \widehat{\boldsymbol{\zeta}}_{1:t-1} \right) \left[\boldsymbol{\Sigma}^\times \right]^{-1} \boldsymbol{\Sigma}^\star \quad (9)$$

The solution (6)-(9) is simply the generalized least squares solution conformed to our present system. For even modest n and τ , this $O[(n\tau)^3]$ solution is, at the very best, computationally impractical.

Consider the real-time, computationally easy approximation in (10)-(15).

$$g_t = (g_{t-1} + \rho) / (g_{t-1} + \rho + 1) \quad (10)$$

$$\mathbf{L}_{\text{HH}t} = \mathbf{L}_{\text{HH}t-1} + g_t \left(\mathbf{H}_t^T \mathbf{H}_t - \mathbf{L}_{\text{HH}t-1} + \mathbf{I}\lambda \right) \quad (11)$$

$$\mathbf{L}_{\text{Hz}t} = \mathbf{L}_{\text{Hz}t-1} + g_t \left(\mathbf{H}_t^T \mathbf{z}_t - \mathbf{L}_{\text{Hz}t-1} \right) \quad (12)$$

$$\mathbf{b}_t = \left[\mathbf{L}_{\text{HH}t-1} \right]^{-1} \mathbf{L}_{\text{Hz}t-1} \quad (13)$$

$$\widehat{\mathbf{z}}_t^T = \mathbf{H}_t \mathbf{b}_{t-q} \quad (14)$$

$$\widetilde{\mathbf{z}}_t = \widehat{\mathbf{z}}_t + \Upsilon_t \quad (15)$$

If $\forall t \in 1 : \tau$, Υ_t is an empty matrix, and hence $\widetilde{\mathbf{z}}_t = \widehat{\mathbf{z}}_t$, our algorithm becomes a fairly generic

variant of ALS [3, 6, 9]. Our ALS process recursively updates the $p \times p$ estimate of the covariate-covariate variance matrix, $\mathbf{L}_{\text{HH}t}$ (11), and our $p \times 1$ estimate of our covariate-response covariance matrix \mathbf{L}_{Hz} (12). The updates are done element-wise, and the amount of the adjustments determined through our scalar gain, g , which is in turn determined by the magnitude of the non-negative hyperparameter, ρ . The hyperparameter λ is our regularizer. The parameter backshift lag, q , will typically take on one of only two possible values set by the researcher: $q \in \{-1, 0\}$. When $q = -1$, we call $\tilde{\mathbf{z}}_t$ the *a priori* predictor of \mathbf{z}_t ; when $q = 0$, we call $\tilde{\mathbf{z}}_t$ the *a posteriori* predictor.

2.2 Interpolation

If we additionally wish to predict to n^* unmonitored locations, we append the preceding ALS algorithm with two more assignments (16) and (17).

$$\hat{\mathbf{z}}_t^{*T} = \mathbf{H}_t^* \mathbf{b}_{t-j} \quad (16)$$

$$\tilde{\mathbf{z}}_t^* = \hat{\mathbf{z}}_t^* + \Upsilon_t^* \quad (17)$$

with $\tilde{\mathbf{z}}_t^*$ being $1 \times n^*$, and the known covariates for the unmonitored sites, \mathbf{H}_t^* , being $n^* \times p$.

3 WIDALS

Supposing that our data arose through (1)-(2), where the measurement errors $\boldsymbol{\varepsilon}_t$ are uncorrelated across time, if we wish to predict $\tilde{\mathbf{z}}_t^*$ given $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_t)$, our immediate inclination might be to uphold the spirit of GLS and have

$$\hat{\boldsymbol{\varepsilon}}^\times = \mathbf{z}_t - \hat{\mathbf{z}}_t \quad (18)$$

$$\mathbf{W}_t^* = [\Sigma_t^\times]^{-1} \Sigma_t^* \quad (19)$$

$$\Upsilon_t = \hat{\boldsymbol{\varepsilon}}^\times \mathbf{W}_t^* \quad (20)$$

where $\Sigma_t^\times = \text{Var}[\mathbf{z}_t]$ and $\Sigma_t^* = \text{Cov}[\mathbf{z}_t, \tilde{\mathbf{z}}_t^*]$. As in kriging, these covariance objects may be modeled through a covariogram, so that $\Sigma_t^\times = C_t(\mathbf{D}^\Omega)$ and $\Sigma_t^* = C_t(\mathbf{D}^{\Omega^*})$. Computationally, the best case is one where the covariances are time-invariant, but nonetheless, unless Σ_t^\times is sparse, inversion is $O[n^3]$. In cases where n is particularly large, say $n > 20000$, even a single inversion may be burdensome. This is compounded by the fact that, despite its small hyperparameter space, ρ , λ , and whatever hyperparameters are required for the covariogram, $C(\cdot)$, fitting our algorithm may require many passes over the data.

Although in a different context, several authors have proposed computationally simplified means of

producing an analogous weighting matrix, \mathbf{W} . Furrer, Genton, and Nychka [2] make use of *tapering* — thresholding covariogram estimates of these two covariance objects by zeroing sufficiently small matrix entries, hence allowing for sparse matrix inversion. As ingenious as this approach is, it brings particular challenges. For it to be practical, large scale unexplained variation in the response surface must be eliminated. For example, if one is working on a spacial domain that is, say, a 100 kilometer square, then we would like to see $C(\cdot)$ vanish for relatively small distances, say distances greater than 20 kilometers or so, in order to really reap computational benefits from space matrix representation. But explaining away all traces of large scale variation either through the inclusion of covariates or spacial bases transforms into \mathbf{H}_t may be non-trivial. Bases transforms in particular may also bring instability in regions of space that are sparsely monitored.

Instead WIDALS makes use of a simple *proxy* function as a replacement for $[\Sigma_t^\times]^{-1} \Sigma_t^\star$. We have

$$\widehat{\boldsymbol{\varepsilon}}^\times = \mathbf{z}_t - \widehat{\mathbf{z}}_t \tag{21}$$

$$\mathbf{W}_t^\star = f_t(\mathbf{D}^{\Omega^\star}) \tag{22}$$

$$\Upsilon_t^\star = \widehat{\boldsymbol{\varepsilon}}^\times \mathbf{W}_t^\star \tag{23}$$

where $f_t(\cdot)$ is trivially some inverse distance function, e.g., $f_t(\mathbf{D}^{\Omega^\star}) = \exp[-\alpha_t \mathbf{D}^{\Omega^\star}]$

3.1 Some WIDALS Extensions & Variants by Example

In the preceding example offering GLS and WIDALS constructions of our stochastic adjustment matrix, Υ_t^\star , we assumed the system measurement errors were temporally uncorrelated. In the subsection that immediately follows, §3.1.1, we will extend (21)-(23) so to accommodate the adjustment of our predictors using multiple lags. In §3.1.2 we'll briefly consider three inverse distance functions.

3.1.1 Construction of Distance in Space-Time

Let's suppose we wish to forecast (to time t) to all monitored locations, ω , one step ahead utilizing lag-1 and lag-2 information in the stochastic adjustment, Υ . We supply $\widehat{\boldsymbol{\varepsilon}}^\times = (\mathbf{z}_{t-2}, \mathbf{z}_{t-1}) - (\widehat{\mathbf{z}}_{t-2}, \widehat{\mathbf{z}}_{t-1})$. We construct distance over $\Omega \times T$ in the following way. Have $d_{i,j}^\Omega$ represent the i, j entry of our $n \times n$ spacial distance matrix, \mathbf{D}^Ω .

$$d_{i,j}^\Omega = \sqrt{\Delta_{x_{i,j}}^2 + \Delta_{y_{i,j}}^2} \tag{24}$$

For some scalar hyperparameter, γ , distance across time is constructed by

$$d_{i,j}^{(\Delta_t)} = \gamma \cdot \Delta_t \quad (25)$$

E.g., $d_{i,j}^{(0)} = 0$, and $d_{i,j}^{(1)} = \gamma$, and $d_{i,j}^{(2)} = 2 \cdot \gamma$, and so on. So that, in our lag $\mathbb{b} = (-2, -1)$ example,

$$\mathbf{D}^* = \sqrt{\begin{pmatrix} \mathbf{D}^{(2)} \\ \mathbf{D}^{(1)} \end{pmatrix}^2 + \begin{pmatrix} \mathbf{D}^\Omega \\ \mathbf{D}^\Omega \end{pmatrix}^2} \quad (26)$$

noting that \mathbf{D}^* is the desired $2n \times n$. Also note that each time distance matrix is constant. To illustrate, suppose we have two monitored sites, $\omega_1 = (1, 2)$ and $\omega_2 = (2, 2)$. So then, using Euclidean distance, we have

$$\mathbf{D}^\Omega = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{D}^{(1)} = \begin{pmatrix} \gamma & \gamma \\ \gamma & \gamma \end{pmatrix} \quad \mathbf{D}^{(2)} = \begin{pmatrix} 2\gamma & 2\gamma \\ 2\gamma & 2\gamma \end{pmatrix} \quad (27)$$

For a second example, suppose we wish to retrospectively predict to a collection of n^* unmonitored locations, ω^* using lags $\mathbb{b} = (-1, 0, 1)$. Then $\widehat{\boldsymbol{\varepsilon}}^\times = (\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1}) - (\widehat{\mathbf{z}}_{t-1}, \widehat{\mathbf{z}}_t, \widehat{\mathbf{z}}_{t+1})$. As in the prior forecast example, the i counter from (24) will index over the n known locations, but j will index over the n^* new locations, ω^* . To illustrate, suppose as above we have two monitored sites at $\omega_1 = (1, 2)$ and $\omega_2 = (2, 2)$. We wish to interpolate to the location $\omega^* = (5, 5)$.

$$\mathbf{D}^\Omega = \begin{pmatrix} \sqrt{4^2 + 3^2} \\ \sqrt{3^2 + 3^2} \end{pmatrix} = \begin{pmatrix} 5 \\ 3\sqrt{2} \end{pmatrix}, \quad \mathbf{D}^{(-1)} = \begin{pmatrix} -\gamma \\ -\gamma \end{pmatrix}, \quad \mathbf{D}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{D}^{(1)} = \begin{pmatrix} \gamma \\ \gamma \end{pmatrix} \quad (28)$$

So ultimately,

$$\mathbf{D}^* = \begin{pmatrix} \sqrt{25 + \gamma^2} \\ \sqrt{18 + \gamma^2} \\ 25 \\ 3\sqrt{2} \\ \sqrt{25 + \gamma^2} \\ \sqrt{18 + \gamma^2} \end{pmatrix} \quad (29)$$

3.1.2 Choices for $f(\cdot)$, the Inverse Distance Function

For the sake of this present work, we will briefly entertain three possibilities for our proxy function.

$$f_t(\mathbf{D}^*) = \exp[-\alpha_t \mathbf{D}^*] \quad (30)$$

And,

$$\mathbf{W}^{\star'} = \exp[-\alpha_t \mathbf{D}^{\star}] \quad (31)$$

$$f_t(\mathbf{D}^{\star}) = \phi \mathbf{W}^{\star'} / \mathbf{w}' \quad (32)$$

where \mathbf{w}' is a vector of column sums of $\mathbf{W}^{\star'}$; the matrix division given in (32) creates columns that sum to one. This is scaled by a flattening hyperparameter, ϕ . Ultimately, then, the columns of $f_t(\mathbf{D}^{\star})$ each sum to ϕ . And finally,

$$\mathbf{D}^{\star'} = \mathbf{D}^{\star} / \bar{\mathbf{d}} \quad (33)$$

$$\mathbf{W}^{\star'} = \exp \left[-\alpha_t \mathbf{D}^{\star'} \right] \quad (34)$$

$$f_t(\mathbf{D}^{\star}) = \phi \mathbf{W}^{\star'} / \mathbf{w}' \quad (35)$$

where $\bar{\mathbf{d}}$ is a vector of length n^{\star} of column means of \mathbf{D}^{\star} . I confess that this “standardizing” of the distances is decidedly *ad hoc*. The effect, obviously, is to shrink the spacial surface around those sites that are relatively remote. While I cannot, at the time of this writing, offer any clear justification for such a transformation, numerous applications of this weighting function to various data, especially those where sites are spacially clustered, produce low RMSE, and also resulted in decidedly attractive interpolation raster maps.

3.2 The WIDALS Hyperparameter Space

In the case where $f_t(\cdot)$ is modeled as time-invariant, the WIDALS solution, all told, requires 5 non-negative scalar hyperparameters, $\Pi^{\text{WID}} = \{\rho, \lambda, \alpha, \gamma, \phi\}$. The first two, ρ and λ are the ALS signal-to-noise ratio (SNR) and regularizer respectively. Distance decay of the mutual influence between sites is determined by α . Increasing alpha has the direct computational effect of removing the influence of distant site response values on the final predicted response value of another site. A measure of distance in time is controlled through γ . Finally, we use ϕ to control the total amount of the adjustment made by the residuals, $\hat{\boldsymbol{\varepsilon}}^{\times}$.

3.3 WIDALS Prediction Quality

In a strictly parametric context, imagining the generative system (1)-(2), where the errors are uncorrelated across time, suboptimality enters our WIDALS process in three places. The first two appear in the deterministic part (i.e., ALS). As discussed in §2, ALS, as described, excludes some response covariation in the production of \mathbf{b} . Moreover, to avoid inverting the error covariance matrix, which is at least $n \times n$, we have excluded it from the process. That is, in (11) we present

$\mathbf{L}_{\text{HH}t}$ as an estimator for $\mathbf{H}_t^T \mathbf{H}_t$ rather than $\mathbf{H}_t^T \Sigma_t^{-1} \mathbf{H}_t$; likewise, Σ_t^{-1} is absent in the creation of $\mathbf{L}_{\text{Hz}t}$. Now, we have introduced suboptimality through the proxy function Υ . A model-specific grand theory concerning the unified contribution of these three error sources in the overall WIDALS prediction error is made particularly challenging since they will be cross-correlated. Presently, at least, we'll consider the sub-optimality of $f_t(\mathbf{D}^*)$ offered in 32.

3.3.1 Miss IDed W

For the following we drop the time domain and suppose we are predicting a single response, \tilde{z}^* . Suppose $\Sigma = C(\mathbf{D}^\Omega)$ is the true measurement-space error variance, and $\mathbf{C} = \hat{C}(\mathbf{D}^\Omega)$ is some estimate of Σ . Have $\sigma^* = C(\mathbf{d}^{\Omega^*})$ and $\mathbf{c}^* = \hat{C}(\mathbf{d}^{\Omega^*})$. The prediction error attributable to the stochastic adjustment, i.e., Υ^* from (17), comes by way of

$$\text{MSE}[\tilde{z}^*] = C(0) - 2 \mathbf{c}^{*T} \mathbf{C}^{-1} \sigma^* + \mathbf{c}^{*T} \mathbf{C}^{-1} \Sigma \mathbf{C}^{-1} \mathbf{c}^* \quad (36)$$

Notice that when $\mathbf{C} = \Sigma$,

$$\text{MSE}[\tilde{z}^*] = C(0) - \mathbf{c}^{*T} \mathbf{C}^{-1} \mathbf{c}^* \quad (37)$$

$$= C(0) - \sigma^{*T} \Sigma^{-1} \sigma^* \quad (38)$$

whose square root is recognized commonly as the *Kriging* prediction error [2, 11].

These results provide for straight forward assessment of a candidate \mathbf{w}^* against some reckoned system covariance, since we may rewrite (36) as

$$\text{MSE}[\tilde{z}^*] = C(0) - 2 \mathbf{w}^{*T} \Sigma^* + \mathbf{w}^{*T} \Sigma \mathbf{w}^* \quad (39)$$

Infill cost (in terms of RMSE) associated with using \mathbf{W} instead of a true matern covariogram, with various parameterizations, is illustrated in Figures 1-3. The leftmost plots in each panel show the covariogram over distance; this distance being roughly the range of the infill region. The middle column plots in each panel show the RMSE of \mathbf{W} divided by the ideal RMSE from (38) against number of sites, n , where we utilize Metropolis-Hastings to locate α and ϕ to produce minimum RMSE for $f(\mathbf{D}^*)$ according to (39). Notice the difference in ordinate scale of these RMSE ratios between the different panels. The rightmost plots in each panel show the change in the RMSE ratio as n increases.

The cost of using our proxy inverse distance function over truth appears minimal when spacial covariation drops rapidly over distances close to zero, that is, when the matern shape parameter is relatively small. In the case where $\kappa = 0.3$, it appears that in all three examples of range, the

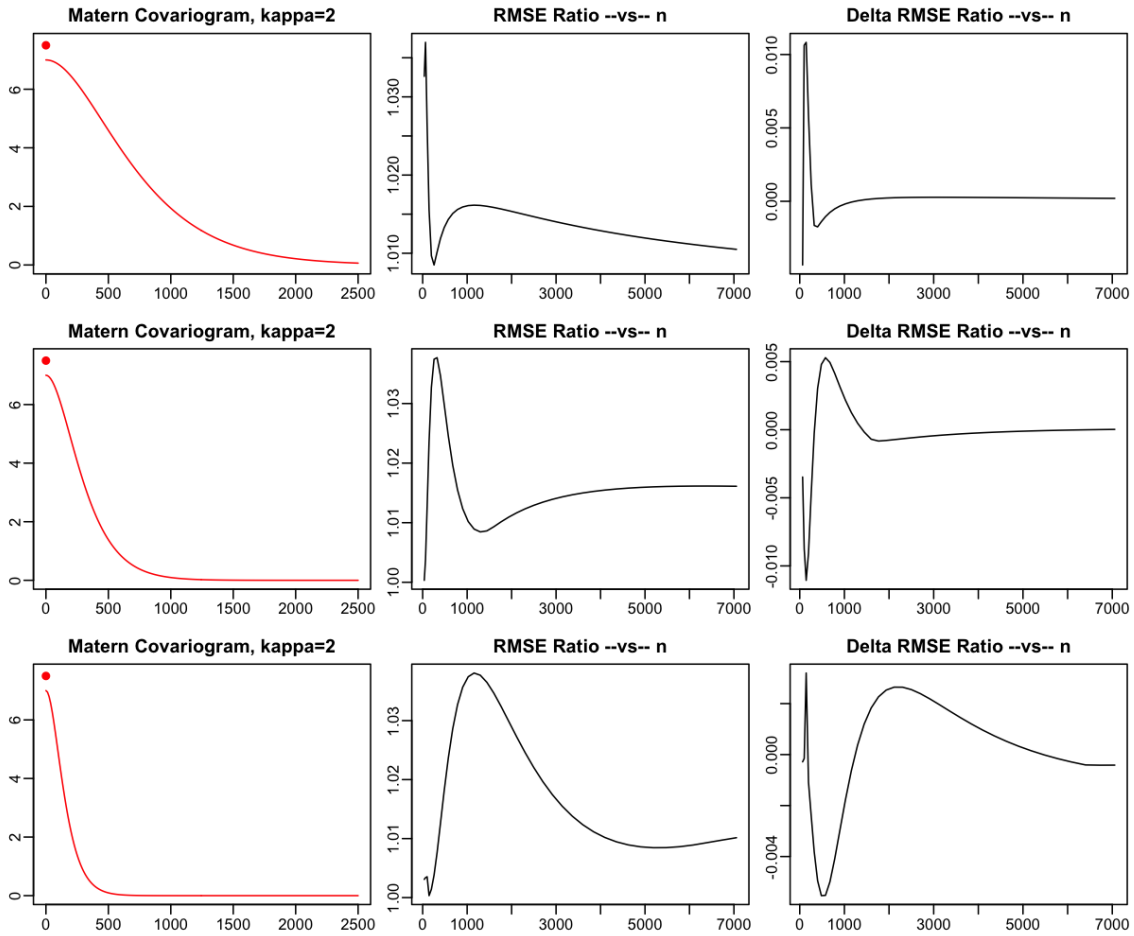


Figure 1: Infill error of miss-specification implied by W for 3 covariogram ranges, $\kappa = 2$, “nugget” effect set to 0.5.

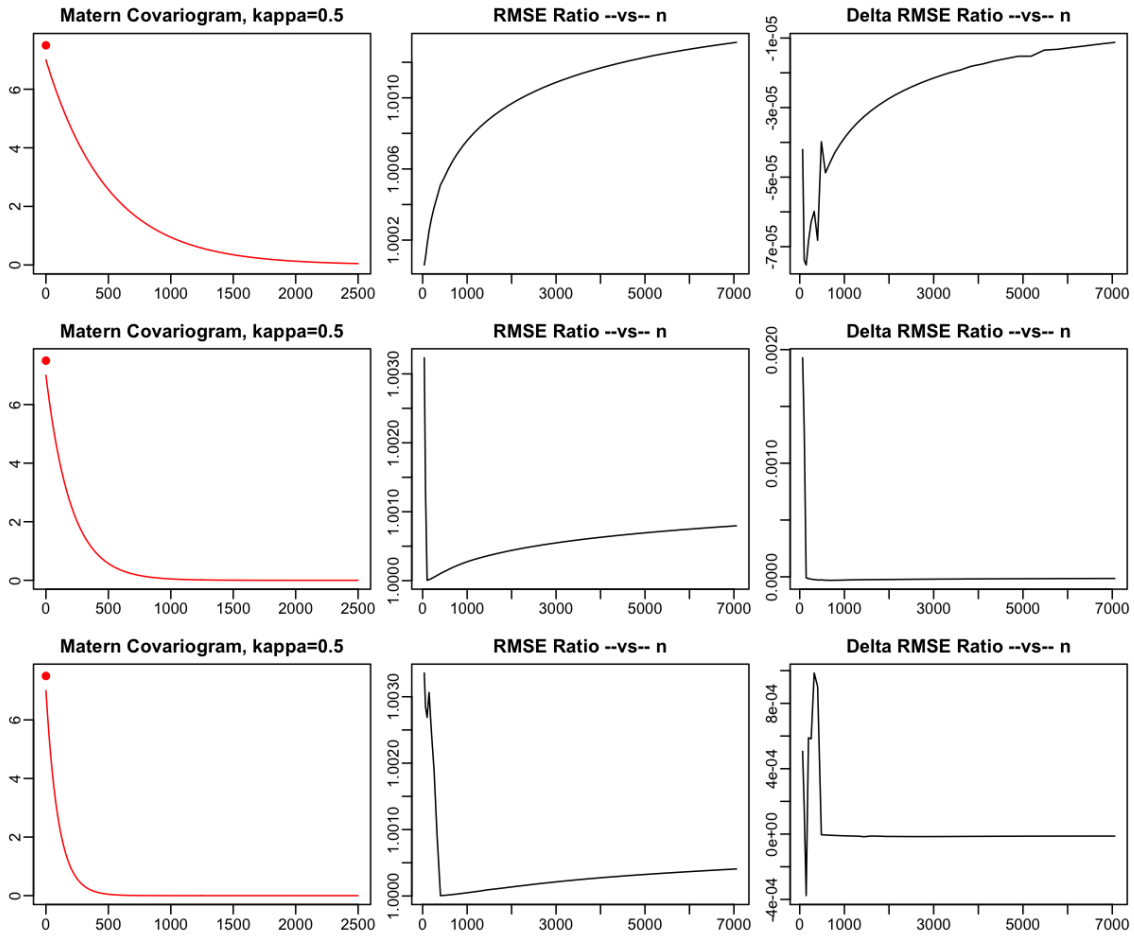


Figure 2: Infill error of miss-specification implied by W for 3 covariogram ranges, $\kappa = 0.5$, “nugget” effect set to 0.5.

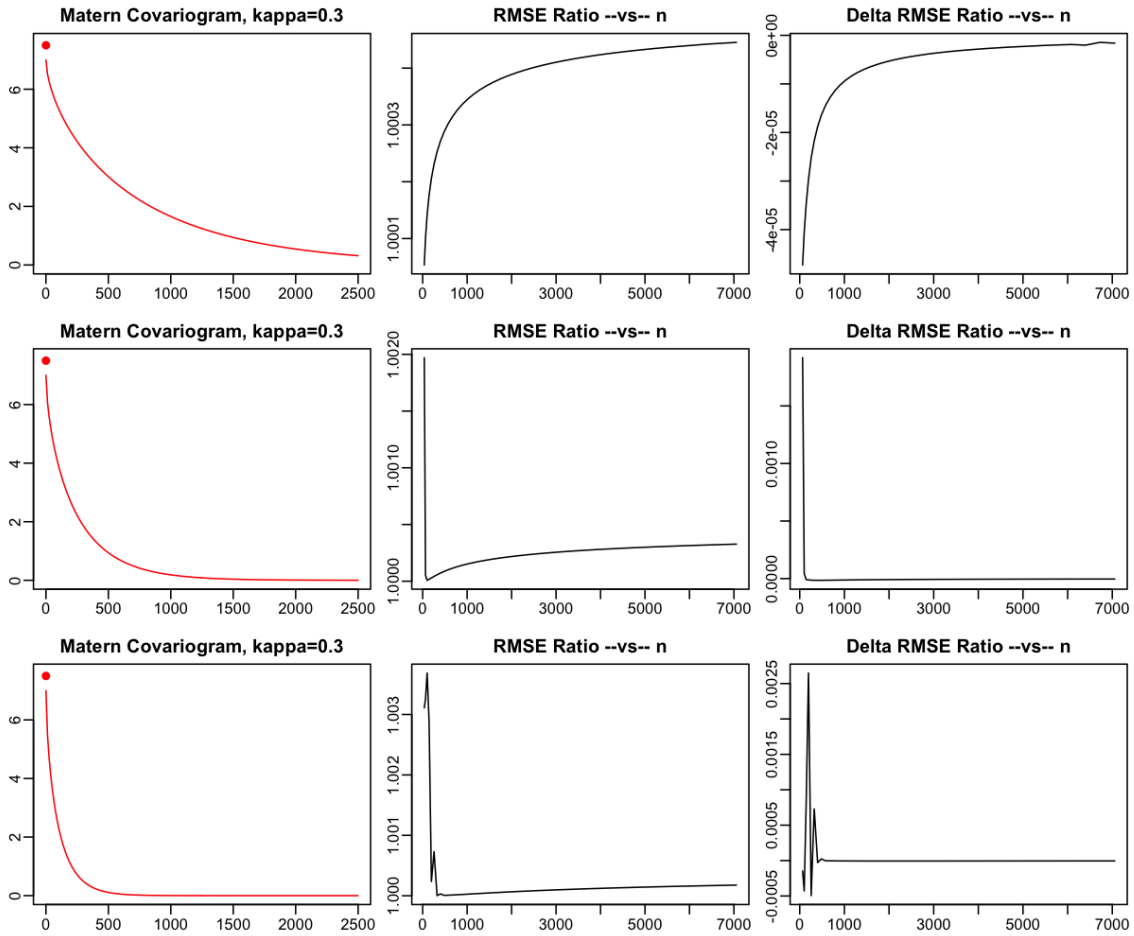


Figure 3: Infill error of miss-specification implied by W for 3 covariogram ranges, $\kappa = 0.3$, “nugget” effect set to 0.5.

RMSE ratio tends to about 1.0003 to 1.0004 for large n . In the worst case presented, where $\kappa = 2$, the ratio tends to about 1.01.

3.4 Fitting: Locating the WIDALS Hyperparameters

For the following demonstration, as well as modeling on numerous simulated and real data, I have found that a simplified version of Metropolis-Hastings (MH) works especially well locating Π^{WID} for the minimization of some cost function, e.g., MSE. Since all five hyperparameters are non-negative, I have the MH stochastic search density be log-normal. Moreover, experience has shown the error surface over the domain of Π^{WID} to be well-behaved, so we can set the MH acceptance probability for a candidate value of Π^{WID} to be degenerate at the previous best.

If we are forecasting to monitored sites, say $\mathbf{b} = (-2, -1)$, we set $\mathbf{q} = -1$, then assessing a hyperparameterization is straight forward, e.g., $\text{MSE} = \text{avg} [\mathbf{Z}_{[\text{train}]} - \tilde{\mathbf{Z}}_{[\text{train}]}]$.

In the case where we wish to interpolate unmonitored responses at time t given $\mathbf{Z}_{[1:t]}$, say $\mathbf{b} = (-1, -0)$, the immediate choice is some form of cross-validation (CV). Suppose $n = 1000$. We could remove $n^* = 100$ sites, set $\mathbf{q} = 0$, then measure a candidate Π^{WID} with $\text{MSE} = \text{avg} [\mathbf{Z}_{[\text{train}]}^* - \tilde{\mathbf{Z}}_{[\text{train}]}^*]$. The obvious shortcoming is that we’re measuring the quality of a set of hyperparameters using the prediction errors of only one-tenth the data. A little more satisfying would be something like 10-fold CV, where the above process of interpolating to 100 removed sites is repeated 10 times, covering all 1000 locations, then examining the resulting MSE across all sites. This, though, would require 10 passes to assess a candidate Π^{WID} .

In lieu of CV, we can take advantage of a subtle property of ALS. It turns out that for large n (precisely the case in which we’re interested), $\mathbf{b}_t \approx \mathbf{b}_{t-1}$.¹ So, we can assess an interpolated fit in the following way. We set $\mathbf{q} = -1$. We then zero-out all the self-referencing weights in \mathbf{W}_t^* . That is, in our current example, the top half of \mathbf{W}_t^* will be an $n \times n$ matrix. We replace the diagonal entries of this matrix with zeros. We do the same for the bottom half of \mathbf{W}_t^* . We dub this method of fitting, pseudo cross-validation (PCV). It happens that the minimum RMSE and the concomitant hyperparameter set gleaned using PCV will often be virtually identical to that wrought with the much slower multi-fold CV.

¹This property can be understood through the intermediate calculation of the so-called “effective sample size.” The reader can consult McCulloch in [6]; I have also detailed a workup in a paper currently undergoing review for publication in *The Journal of Environmental Statistics*.

4 Application of WIDALS to Global Daily Temperature

We obtained daily high temperature, in degrees Fahrenheit, for $n = 5182$ sites across the globe, for the dates May 2, 2010 to June 1, 2011, for a total of $\tau = 396$ epochs [7].

We created 4 covariates. Two are time-invariant. The first and second columns of \mathbf{H}_t contain a constant term and location elevation. The other two are both space- and time-variant — they are *non-seperable* in space-time. The third column of \mathbf{H}_t contains the noontime solar energy incident to the site (40)-(42).

$$\psi = 23.5 \sin[\delta 360/365.25] \quad (40)$$

$$\eta = 90 - \arccos[\cos[y] \cos[\psi] + \sin[\psi] \sin[y]] \quad (41)$$

$$A(\delta, y) = \sin[\eta] \quad (42)$$

where δ is the number of days past the Spring equinox (March 20th), and where y is latitude in degrees. The fourth column of \mathbf{H}_t contains the interaction between this incident solar energy and elevation.

We construct \mathbf{D}^Ω using global geodesic distance in units kilometers.

We construct our model to interpolate to time t by setting $\mathbf{b} = (-1, 0)$, $\mathbf{q} = -1$. We construct a time-invariant Υ using (33)-(35). We simultaneously train and test over $t \in 25 : 394$ using MH on the PCV method described in the previous section. We obtain an RMSE = 3.587 with $\Pi^{\text{WID}} = \{\rho, \lambda, \alpha, \gamma, \phi\} = (1.1684 \cdot 10^{-8}, 4.041 \cdot 10^{-5}, 63.1, 410, 1.0040)$

parameter est.	value	als.se
Intercept	4.8119	0.0338
Elevation	-8.6408	0.0293
Sun Energy	71.9304	0.0450
Elevation * Sun Energy	6.4704	0.0369

Table 2: Value of ALS parameter, \mathbf{b}_{396} , and approximate standard errors.

We then construct a vector of 41,437 approximately evenly spaced new sites covering the globe, ω^* . We acquired elevations for these sites and computed each site's energy of solar incidence to construct \mathbf{H}_t^* .

The resulting interpolated WIDALS predictions are shown in the following video.

*** Global Daily Max Temperature ***

If we use only ALS for interpolation, that is, if Υ is empty, we get an RMSE = 11.04.

References

- [1] Stephen D. Billings, Rick K. Beatson, and Garry N. Newsam. Interpolation of geophysical data using continuous global surfaces. *Geophysics*, 67(6):1810–1822, 2002.
- [2] Reinhard Furrer, Marc G. Genton, and Douglas Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15:502–523, 2006.
- [3] Simon Haykin. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, forth edition, 2002.
- [4] Pierre Delfiner Jean-Paul Chilès. *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons, New York, NY, 1999.
- [5] Phaedon C. Kyriakidis and André G. Journel. Geostatistical space-time models: A review. *Mathematical Geology*, 31, 1999.
- [6] J. Huston McCulloch. The kalman foundations of adaptive least squares, with application to u.s. inflation. *Unpublished*, 2005.
- [7] National Climatic Data Center. <http://www.ncdc.noaa.gov/oa/climate/climatedata.html#surface>. —.
- [8] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.
- [9] Ali H. Sayed. *Fundamentals of Adaptive Filtering*. John Wiley & Sons, Hoboken, N.J., 2003.
- [10] Michael L. Stein. The screening effect in kriging. *The Annals of Statistics*, 30(1):298–323, 2002.
- [11] Michael L. Stein and Mark S. Handcock. Some asymptotic properties of kriging when the covariance function is misspecified. *Mathematical Geology*, 21(2), 1989.
- [12] R. Van der Merwe and E.A. Wan. The square-root unscented kalman filter for state and parameter-estimation. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 6, pages 3461–3464, 2001.