

— Under Review —

MMMMMM YYYY, Volume VV, Issue II.

<http://www.jenvstat.org>

Facile Spacio-Temporal Modeling

Forecasting with Adaptive Least Squares and the Kalman Filter

Dave Zes

Department of Statistics, UCLA

Abstract

In the following we examine, compare, and to a point, advocate simple methods of spacio-temporal description and forecasting. Included are the two-level state-space system, and time-varying parameter least squares auto-regressive system, along with their respective solving algorithms, the Kalman Filter, and auto-regressive Adaptive Least Squares (ALS). Advantages especially attributed to ALS include computational frugality, ease of implementation and interpretation, broad applicability, flexibility, and often excellent performance. Additionally, since ALS relies on estimating response covariation, it may also serve as a precursor to space-time interpolation where specification of a covariogram is required. Comparisons on several simulated datasets and three real datasets included.

Keywords: kriging, adaptive least squares, covariance function, covariogram, kalman filter.

1. Introduction

Within the discipline of statistics, the particular sub-discipline that focuses on observations over the space-time domain has recently received considerable interest — the broad range of application, along with the perceived import of these applications being major motivators. Spacio-Temporal estimation provides answers in many different arenas, such as ecology (pollution, new growth, habitat preservation), industry (resource availability, regulation attainment), climate (in every imaginable way), medicine (epidemiology), even in marketing (demographics). There exists a visceral appeal, since when “space” refers to 3-dimensional Euclidean space, the “space-time” domain is the very domain of human experience.

One might view spacial modeling approaches as falling into two camps. In one, somewhat in the spirit of ordinary regression, we emphasize a smooth function, $y(\cdot)$, over space that passes in some meaningful way *through* the observations, \mathbf{z} . One symbolic interpretation of this might look something like $\boldsymbol{\omega} \in \Omega \subset \mathbb{R}^2$, $\mathbf{y} = f(\boldsymbol{\omega})$, and $\mathbf{z}(\boldsymbol{\omega}) = \mathbf{y}(\boldsymbol{\omega}) + \boldsymbol{\varepsilon}$, where

$\boldsymbol{\omega}$ is a vector of n locations, each location being defined by its coordinates, $\boldsymbol{\varepsilon}$ is a vector of independent zero-mean shocks. The other camp offers that what might appear as local “structure” in the observations does not result from some deterministic mean function, but rather from a stochastic process that favors covariance between nearby observations: $\boldsymbol{z} = \boldsymbol{\delta}$, with $\boldsymbol{\delta} \sim \mathcal{N}[\mathbf{0}, \Sigma]$, where it is commonly further assumed that the covariance, Σ , is equal to a covariance function, $C(\cdot)$, operating on the distance matrix, \mathbf{D} , between our site locations, $\boldsymbol{\omega}$. Both Gneiting, Genton, and Guttorp (2005) and also Kyriakidis and Journel (1999) offer nice exposition surveying some popular methods. In both camps, the domain, Ω , is utilized, though in different ways. In the former, the mean function, $y(\Omega)$, might be taken to be some transform of Ω . In the latter, as with ordinary kriging, for example, we use only metric features of Ω to estimate the covariance structure.

As is virtually always the case in applied modeling, assumptions at some point fall back on a mathematic idealization of nature; however, a preference towards one view over the other — believing that an observation at some location $z(\omega_0)$ acquires its value from nearby locations rather than properties, either latent or observable, of the spacial domain — can be philosophically provocative. The position that “the temperature at a location in Death Valley will more strongly covary with a nearby measurement than one taken at the Port of Los Angeles not because of *distance*, but rather because there are different climatological processes at play at the beach than at the desert,” is just as sound as the statement “spacial covariation in temperature results from the diffusion of heat.”

In practice, researchers commonly take a complementary approach and offer that the phenomenon under study arises through something akin to $\boldsymbol{z}(\boldsymbol{\omega}) = \boldsymbol{y}(\boldsymbol{\omega}) + \boldsymbol{\delta}$, with $\boldsymbol{\delta} \sim f(\mathbf{0}, \Sigma)$, where local covariation and spacial structure each provide a discrete, arithmetic contribution to the response. In fact, the spacio-temporal literature is dominated by such assumed systems — or the more generalized temporal adaptation offered by the two-level *state-space* system (1)-(2). Here, the introduction of a time dimension — discrete and evenly spaced in typical applications, $t \in T = 1 : \tau$ — extends our spacial model in much the way we might expect. The mean function may be allowed to evolve, either deterministically or stochastically, over time.

Adaptive Least Squares (ALS) is close kin to Least Mean Squares (LMS) and Recursive Least Squares (RLS), techniques that, much like the Kalman Filter, had seminal days back in the 1960s. While we offer some modestly novel theoretical exposition in Section 4, and some novel applications in Section 7, these techniques have been widely explored. A great majority of this literature, though, has been authored by, and for, engineers. The primary goal of this present work is to reinvigorate, for a statistical audience, especially those keen on space-time modeling, interest in ALS. When forecasting to monitored locations, ALS is an extremely powerful algorithm, possessing several virtues, e.g., requiring no explicit knowledge of $y(\Omega)$ or Σ , and also being widely extensible. In addition, as we will see in some correlation videos, it can also serve as a prelude to covariogram analysis for the sake of spacial interpolation (predicting to an unmonitored location).

1.1. Some Notation, Terminology

Generally, lowercase bold refers to a vector, uppercase bold, a matrix. Our longitudinal data matrix, \mathbf{Z} , has rows indexed by time, so that extracting \boldsymbol{z}_t from \mathbf{Z} produces a row vector. Where appropriate, we will enforce the notation that an italic, or leaning character, indicates

a random variable, whereas an upright, or roman-style character, will refer to a constant. We use $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_n) \in \Omega \subset \mathbb{R}^m$, where often in application, $m = 2$, to refer to our n spacial locations. We introduce a discrete time dimension, $t \in T$, and use $z_t(\omega_i)$, or just $z_{t,i}$, to refer to the random response variable at site ω_i at time t , and $\mathbf{z}_t(\boldsymbol{\omega})$, or just \mathbf{z}_t , to refer to our $1 \times n$ observation vector. For the sake of compactness, $\mathbf{z}_t(-i)$, equivalent to $\mathbf{Z}_{[t,-i]}$, refers to the row vector of observations at time t with the observation at site i removed. A “wiggly” spacial mean function will be one whose mean second derivative magnitude is relatively large, sometimes quantified at large using $\int_{\Omega} (y''(\Omega))^2 d\Omega$. Lastly, it should be noted that the three particular variants of ALS that we will entertain in the following are all members of the class of Auto-Regressive Adaptive Least Squares, that we will often call by default, simply, “ALS.”

2. The Kalman Filter

Consider the following spacio-temporal model.

$$\boldsymbol{\beta}_t = \mathbf{F} \boldsymbol{\beta}_{t-1} + \boldsymbol{\nu}_t \quad (1)$$

$$\mathbf{z}_t^T = \mathbf{H} \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \quad (2)$$

with $\boldsymbol{\nu}_t \sim \mathcal{N}[\mathbf{0}, \mathbf{Q}]$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}[\mathbf{0}, \mathbf{R}]$. From an engineering perspective, there exist two exhaustive and authoritative texts covering the larger class of dynamic systems, [Sayed \(2003\)](#) and [Haykin \(2002\)](#). Both are comprehensive and desirable in every didactic way. Most important statements presented here on this class of systems can be found in either text. From a statistics perspective, the equally comprehensive and desirable [West and Harrison \(1997\)](#) can be consulted, while [Shumway and Stoffer \(2006\)](#) offer a compact discussion. In engineering, (1) and (2) describe a type of “state-space” system. In statistics, this system is known as a type of “time-varying parameters” model. These names can be reconciled. Notice that this system looks much like a regression model where \mathbf{H} plays the role of design matrix, and $\boldsymbol{\beta}_t$, the partial slopes — or the model *parameters*, which are indeed time-varying.

The process of estimating the system hyperparameters, $\{\mathbf{F}, \mathbf{H}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\beta}_0\}$, in an assumed functional relationship from historical data, \mathbf{Z} , may be called “grey box system identification” to engineers, or generically as “model fitting” to the statistician. The first level in the system’s recursive hierarchy, (1), is known as the “state equation,” the second level, (2), is known in the engineering literature as the “measurement equation,” or “observation equation.”

This system may readily be cast into the spacio-temporal domain. If we call \mathbf{D} the $n \times n$ matrix of distances between $\boldsymbol{\omega}$, then we can account for local spacial covariation by having $\mathbf{R} = C(\mathbf{D})$, where $C(\cdot)$ is some covariogram of our choosing. Moreover, we may have our measurement function, \mathbf{H} , include some bases transform over space, $\mathbf{H} = \mathbf{H}(\boldsymbol{\omega}, \cdot)$.

When all the system shocks and measurement shocks $\boldsymbol{\nu}_t$ and $\boldsymbol{\varepsilon}_t$ from (1)-(2) are uncorrelated, the optimal *unbiased* $L2$ predictor (BLUP) for the latent state, $\boldsymbol{\beta}_t$, comes by way of the

recursive process

$$\widehat{\boldsymbol{\beta}}_t^{(-)} = \mathbf{F}\widehat{\boldsymbol{\beta}}_{t-1}^{(+)} \quad (3)$$

$$\mathbf{P}_t^{(-)} = \mathbf{F}\mathbf{P}_{t-1}^{(+)}\mathbf{F}^T + \mathbf{Q} \quad (4)$$

$$\mathbf{K}_t = \mathbf{P}_t^{(-)}\mathbf{H}^T (\mathbf{H}\mathbf{P}_t^{(-)}\mathbf{H}^T + \mathbf{R})^{-1} \quad (5)$$

$$\mathbf{P}_t^{(+)} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_t^{(-)} \quad (6)$$

$$\widehat{\boldsymbol{\beta}}_t^{(+)} = \widehat{\boldsymbol{\beta}}_t^{(-)} + \mathbf{K}_t (\mathbf{z}_t^T - \mathbf{H}\widehat{\boldsymbol{\beta}}_{t-1}^{(-)}) \quad (7)$$

an example of the Kalman Filter (KF). The symbol, $\widehat{\boldsymbol{\beta}}_t^{(-)}$, called the *a priori* predictor, is our prediction of the random variable $\boldsymbol{\beta}_t$ prior to observing \mathbf{z}_t . The symbol $\widehat{\boldsymbol{\beta}}_t^{(+)}$, called the *a posteriori* predictor, is our prediction of the random variable $\boldsymbol{\beta}_t$ after observing \mathbf{z}_t . Since \mathbf{H} is linear, the best unbiased $L2$ predictor of a yet unmeasured \mathbf{z}_t is $\mathbf{y}_t = \mathbf{H}\widehat{\boldsymbol{\beta}}_t^{(-)}$. Or equivalently,

$$\mathbb{E}[\mathbf{z}_t^T \mid \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots, \mathbf{z}_1] = \mathbf{H}\widehat{\boldsymbol{\beta}}_t^{(-)} \quad (8)$$

Notice that we have used random variables to describe the process (3)-(7), i.e., we use \mathbf{z}_t and $\widehat{\boldsymbol{\beta}}_t$ rather than \mathbf{z}_t and $\widehat{\boldsymbol{\beta}}_t$ respectively. This is done to highlight the fact that the value of the error matrix, \mathbf{P}_t , is indifferent to whatever particular data realization \mathbf{Z} of \mathbf{Z} we process. Centrally,

$$\mathbf{P}_t^{(-)} = \mathbb{E} \left[\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(-)} \right) \left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(-)} \right)^T \right] \quad (9)$$

and

$$\mathbf{P}_t^{(+)} = \mathbb{E} \left[\left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(+)} \right) \left(\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t^{(+)} \right)^T \right] \quad (10)$$

measure the expected quality of our *a priori* and *a posteriori* predictions respectively. The inverse, \mathbf{P}_t^{-1} , may rightly be called the system's (or solution's) "precision" matrix.

A "stable" system is one where the system function, \mathbf{F} , lies inside the unit circle, i.e., the magnitude of the largest eigenvalue of \mathbf{F} is less than 1.

Initial conditions: For a stable system, we should set $\boldsymbol{\beta}_0 = \mathbb{E}[\boldsymbol{\beta}] = 0$, and usual convention is to have $\mathbf{P}_0^{(+)} = \mathbf{Q}(\mathbf{I} - \mathbf{F}^T\mathbf{F})^{-1}$. Attention is given in the engineering literature to — what most statisticians would consider the rather unrealistic case of — tracking where a researcher must ready an unstable process in the absence of even a single observation. In such cases one can be highly conservative and have $\mathbf{P}_0 = \text{diag}[\mathbf{P}]$ for some very large value, $\mathbf{P} = 10^{10}$, for example. This is known as imposing a *diffuse*, or *uninformative* initial condition, or, in Bayesian parlance, a *diffuse*, or *uninformative prior*.

3. Adaptive Least Squares

Consider an auto-regressive system independent of Ω , and only dependent on time labels signaling $\Delta t = 1$.

$$\mathbf{z}_t = B(\mathbf{Z}_{1:t-1}) + \boldsymbol{\varepsilon}_t \quad (11)$$

If $B(\cdot)$ is linear, then the optimal forecasting estimator comes by way of

$$\hat{\mathbf{z}}_t = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{t-1}) \mathbb{E}[\Phi_{1:t-1}^T \Phi_{1:t-1}]^{-1} \mathbb{E}[\Phi_{1:t-1}^T \mathbf{z}_t] \quad (12)$$

with

$$\Phi_{1:t-1} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{t-1}) \quad (13)$$

The challenge, of course, is estimating the covariance across lag-0 (the regressor-regressor covariance, the first expectation in (12)) and the covariance between lag-0 and lag-1 (the regressor-response covariance, the second expectation).

We now consider a special AR1 case of (11):

$$\mathbf{B}_t = \mathbf{I}\mathbf{B}_{t-1} + \boldsymbol{\nu}_t \quad (14)$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} \mathbf{B}_t + \boldsymbol{\varepsilon}_t \quad (15)$$

with $\boldsymbol{\nu}_t \sim \mathcal{N}[\mathbf{0}, \sigma_\nu \mathbf{I}]$ and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}[\mathbf{0}, \sigma_\varepsilon \mathbf{I}]$. This AR1 variant can be cast as a state-space system, and can be solved utilizing an auto-regressive variant of Adaptive Least Squares (ALS):

$$g_t = (g_{t-1} + \rho) / (g_{t-1} + \rho + 1) \quad (16)$$

$$\hat{\mathbf{B}}_t = \left(\mathbf{L}_{t-1}^{(0)} \right)^{-1} \mathbf{L}_{t-1}^{(1)} \quad (17)$$

$$\hat{\mathbf{z}}_t = \mathbf{z}_{t-1} \hat{\mathbf{B}}_t \quad (18)$$

$$\mathbf{L}_t^{(0)} = \mathbf{L}_{t-1}^{(0)} + g_t \left(\mathbf{z}_{t-1}^T \mathbf{z}_{t-1} - \mathbf{L}_{t-1}^{(0)} + \mathbf{I}\lambda \right) \quad (19)$$

$$\mathbf{L}_t^{(1)} = \mathbf{L}_{t-1}^{(1)} + g_t \left(\mathbf{z}_{t-1}^T \mathbf{z}_t - \mathbf{L}_{t-1}^{(1)} \right) \quad (20)$$

with $\rho = \sigma_\nu / (\sigma_\nu + \sigma_\varepsilon)$, and where it is common to have $\mathbf{L}_0^{(0)} = \mathbf{0}$, $\mathbf{L}_t^{(1)} = \mathbf{0}$, and $g_0 = \infty$ (so that $g_1 = 1$). ALS (16)-(20) is a *stochastic descent* operation. ‘‘Stochastic’’ since the covariances, $\mathbf{L}^{(0)}$ and $\mathbf{L}^{(1)}$, are empirically estimated from noisy observations, ‘‘descent’’ since as we collect more and more data as t increases, the objective function decreases. For an attractive introduction into ALS and some filter variants, see McCulloch (2005). The superscript in $\mathbf{L}_t^{(0)}$ indicates that we are interested in an estimate of system covariance of observations separated by 0 time epochs. In (19)-(20) each of the two covariance estimates are recursively updated with uniform adjustments to each element determined by the scalar gain, here notated as convention dictates with g (rather than K). It so happens that our gain update (16) is equivalent to that given previously in (5) when the state is scalar, where $\rho = \sigma_\nu / (\sigma_\nu + \sigma_\varepsilon)$. This Kalman-like gain distinguishes this ALS method from the otherwise similar usual formulations of Recursive Least Squares (RLS), which instead updates covariance estimates using a constant *forgetting factor*. The most well-known version of RLS possesses the potentially desirable property that it does not require inversion of the regressor-regressor variance matrix, $\mathbf{L}_t^{(0)}$. Rather, it makes use of a Sherman-Woodbury-Morrison (SWM) update. For large datasets, say where the number of sites exceeds a few hundred, RLS offers an attractive alternative to (16)-(20). In its typical embodiment, though, RLS lacks regularization. Our regularizer, appearing as $\mathbf{I}\lambda$ in (19), is analogous to *ridge* regularization common to regression. While the regularizer introduces bias in the prediction, it often reduces total prediction error. The importance of this property has inspired clever investigations into uniting it with the

computationally advantageous SWM RLS, Gunnarsson (1996), Gunnarsson (1994), Leung, Young, Sum, and Kan (1999). There exist numerous variants of ALS, the most important ones focusing on manipulating the effect of this regularization. In (19) the effect of *diagonally loading* the predictor-predictor variance matrix, $\mathbf{L}_t^{(0)}$, with $\mathbf{I}\lambda$ is to shrink our predictions, $\hat{\mathbf{z}}_t$, to zero. In cases where the marginal variance of the system is large (and hence the response values drift away from zero), or if the mean function is wiggly (the response values are greatly varied depending upon location), pulling predictions towards zero may be a poor strategy for reducing prediction error. What we can do in such cases is *time-center* the covariance estimates. The effect is to shrink predictions not to zero, but rather towards response values seen in temporally local observations. This centered variant will be detailed in Section 5.

We may generalize our algorithm, (16)-(20), by including more lagged observations. Since the predictor and predictand force the choice of covariances to be estimated, we can write the more general form as

$$g_t = (g_{t-1} + \rho) / (g_{t-1} + \rho + 1) \quad (21)$$

$$\hat{\mathbf{z}}_t = \mathbf{z}^* \left(\mathbf{L}_{t-1}^\times \right)^{-1} \mathbf{L}_{t-1}^* \quad (22)$$

$$\mathbf{L}_t^\times = \mathbf{L}_{t-1}^\times + g_t \left(\mathbf{z}^{*T} \mathbf{z}^* - \mathbf{L}_{t-1}^\times + \mathbf{I}\lambda \right) \quad (23)$$

$$\mathbf{L}_t^* = \mathbf{L}_{t-1}^* + g_t \left(\mathbf{z}^{*T} \mathbf{z}_t - \mathbf{L}_{t-1}^* \right) \quad (24)$$

The predictor, for example, may be $\mathbf{z}^* = (\mathbf{z}_{t-3}, \mathbf{z}_{t-2}, \mathbf{z}_{t-1})$, or, as we will see in our application to atmospheric methane in Section 7.3, if a smoother is sought for imputation, $\mathbf{z}^* = (\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1})$. In this general form the only thing open to decision is what observations are used to estimate the covariances. This decision only occurs when smoothing; in forecasting, our hands are tied. If forecasting to t , we are only permitted use of observations prior to t .

The appeal of ALS is its great parsimony. We can run ALS with as few as a single, non-negative scalar hyperparameter, ρ . But how good a proxy is the time-varying parameter AR1 system (14)-(15) to the state-space system (1)-(2) — or, to the point, under what conditions will ALS compete with an optimal Kalman Filter?

4. Relating Adaptive Least Squares and the Kalman Filter

For a stable system, the rudimentary connection between ALS and the KF as applied to the presently considered state-space system, (1) and (2), is most easily revealed through two equalities.

$$\mathbf{\Lambda}^{(0)} := \mathbb{E}[\mathbf{z}_t^T \mathbf{z}_t] = \mathbf{H} \mathbf{Q} (\mathbf{I} - \mathbf{F}^T \mathbf{F})^{-1} \mathbf{H}^T + \mathbf{R} \quad (25)$$

$$\mathbf{\Lambda}^{(k)} := \mathbb{E}[\mathbf{z}_t^T \mathbf{z}_{t-k}] = \mathbf{H} \mathbf{Q} \mathbf{F}^k (\mathbf{I} - \mathbf{F}^T \mathbf{F})^{-1} \mathbf{H}^T \quad (26)$$

with $k \in (1, 2, \dots, t-1)$. Keeping our sights fixed on 1-step-ahead forecasting to time t , we can rewrite (12) to obtain

$$\mathbf{\Lambda}^\times = \begin{pmatrix} \mathbf{\Lambda}^{(0)} & \mathbf{\Lambda}^{(1)} & \dots & \mathbf{\Lambda}^{(t-2)} \\ \mathbf{\Lambda}^{(1)} & \mathbf{\Lambda}^{(0)} & \dots & \mathbf{\Lambda}^{(t-3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Lambda}^{(t-2)} & \mathbf{\Lambda}^{(t-3)} & \dots & \mathbf{\Lambda}^{(0)} \end{pmatrix}, \quad \mathbf{\Lambda}^* = \begin{pmatrix} \mathbf{\Lambda}^{(t-1)} \\ \mathbf{\Lambda}^{(t-2)} \\ \vdots \\ \mathbf{\Lambda}^{(1)} \end{pmatrix} \quad (27)$$

Finally, we solve

$$\tilde{\mathbf{z}}_t = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{t-1}) (\mathbf{\Lambda}^\times)^{-1} \mathbf{\Lambda}^* \quad (28)$$

and arrive at the following assertion.

Statement 1: The solution, $\tilde{\mathbf{z}}_t$, given in (28), is the Kalman solution.

Our supporting argument takes nary a lick of algebra.

Pf: Since the joint distribution of any arbitrary subsets of observations is Gaussian, by the Gauss-Markov theorem we have that $\tilde{\mathbf{z}}_t = \mathbb{E}[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots, \mathbf{z}_1, \mathbf{H}, \mathbf{Q}, \mathbf{F}, \mathbf{R}]$. We also have the oft-published equality for the Kalman Solution $\hat{\mathbf{z}}_t = \mathbb{E}[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \dots, \mathbf{z}_1, \mathbf{H}, \mathbf{Q}, \mathbf{F}, \mathbf{R}]$. Since the expectation of a random variable is unique, it must be the case that $\tilde{\mathbf{z}}_t = \hat{\mathbf{z}}_t$.

This implies that we can make a twofold distinction between the two solutions. First, in ALS, we use \mathbf{L}^\times and \mathbf{L}^* as proxies for $\mathbf{\Lambda}^\times$ and $\mathbf{\Lambda}^*$ respectively, but only through a usually small number of pre-specified lags (e.g., 1, 2, 3). Second, because we only estimate these covariances across a small number of lags, we necessarily only use a small number of time-lagged observations as our predictors.

We can gain some insight into the difference between the quality of the ALS and Kalman predictions using a single lag simplification of (28) as an intermediate.

$$\tilde{\mathbf{z}}_t = \mathbf{z}_{t-1} (\mathbf{\Lambda}^{(0)})^{-1} \mathbf{\Lambda}^{(1)} \quad (29)$$

where the covariances are defined as in (25) and (26). Notice that this solution is entirely deterministic; both covariances are fixed, and the predictors, once recorded, are treated as fixed.

We can use the fundamental least squares relation $\sigma^2 = \mathbb{E}[(z - \mu_z)^2] - \mathbb{E}[(\hat{z} - \mu_z)^2]$, since $\hat{z} \perp z - \hat{z}$, to derive the 1-step-ahead forecast error. First, define

$$\mathbf{V} := (\mathbf{\Lambda}^{(0)})^{-1} \mathbf{\Lambda}^{(1)} \quad (30)$$

our matrix of slopes that map \mathbf{z}_{t-1} to \mathbf{z}_t , for example, $\hat{z}_{i,t} = \mathbf{z}_{t-1} \mathbf{v}_{\cdot,i}$, where $\mathbf{v}_{\cdot,i}$ is the i th column of \mathbf{V} . Notice that since $\mathbb{E}[\mathbf{z}_t^T \mathbf{z}_t] = \mathbb{E}[\mathbf{z}_{t-1}^T \mathbf{z}_{t-1}]$, \mathbf{V} also serves as the matrix of correlations between \mathbf{z}_t and \mathbf{z}_{t-1} . Let e_i be the 1-step-ahead forecast error at site i . Our fitted variances for site i is

$$\Upsilon_i = \mathbf{v}_{\cdot,i} \mathbf{v}_{\cdot,i}^T \circ \mathbf{\Lambda}^{(0)} \quad (31)$$

where \circ represents the element-wise product. If we have $\mathbf{1}$ denote a column vector of 1's, we arrive at

$$\mathbb{E}[e_i^2] = \mathbf{\Lambda}_{[i,i]}^{(0)} - \mathbf{1}^T \Upsilon_i \mathbf{1} \quad (32)$$

This allows us to compare $\mathbb{E}[e_i^2]$ with, say, the steady-state value of $\mathbf{P}_{[i,i]}^{(-)}$ to provide an assessment of how using only lag-1 observations degrades $\tilde{\mathbf{z}}$ from (29) compared to the Kalman prediction.

In a more general setting, when the state shocks or measurement shocks are cross-correlated across time, e.g., in (14), $\boldsymbol{\nu}_t \not\perp \boldsymbol{\nu}_{t-1}$, or in (15), $\boldsymbol{\varepsilon}_t \not\perp \boldsymbol{\varepsilon}_{t-1}$, determining the theoretical cost of estimating the fully informative $\boldsymbol{\Lambda}^\times$ and $\boldsymbol{\Lambda}^*$ from (27) through the time-censored proxies, \mathbf{L}^\times and \mathbf{L}^* from (23)-(24), is nontrivial. In essence, we use a *temporal neighborhood* of observations to mimic the full covariance structure of the system. Justification, heuristically, rests on the argument that covariation between temporally nearby observations will dominate, or “screen” the effect of temporally distant observations. The spacial screening analogue is explored in Stein (2002); additionally, both Sayed (2003) (pg. 763) and Ljung (1998) (pg. 367) offer lucid workups relating RLS to a highly special case of the Kalman Filter.

Fortunately, there does exist an interpretation that is both clean and fruitful, albeit purely empirical. Let us imagine that the scalar ρ has been set to zero. Then the gain sequence, (g_1, g_2, g_3, \dots) becomes the harmonic sequence $(1, \frac{1}{2}, \frac{1}{3}, \dots)$. At any time, t , the partial slope estimate, $\hat{\mathbf{B}}_t = (\mathbf{L}_{t-1}^\times)^{-1} \mathbf{L}_{t-1}^*$, is simply an evenly-weighted least squares estimate. The so-called “effective sample size” of the process at time t , then, is $g_t^{-1} = t$. This same reasoning can be extended for any value of ρ . If, say, $g_t = 0.01$, then the observed covariances at time t provide a weighted adjustment of 1/100 towards the current estimates of \mathbf{L}^\times and \mathbf{L}^* , and so the effective sample size at time t is 100. We may then calculate the “effective standard errors” of our solution. Have $\mathbf{l}_{t,i}^*$ be the i th column of \mathbf{L}_t^* , and define

$$\mathbf{b}_{t,i} := (\mathbf{L}_t^\times)^{-1} \mathbf{l}_{t,i}^*, \quad \boldsymbol{\Pi} := (\mathbf{L}_t^\times)^{-1}, \quad \hat{\sigma}_{t,i}^2 := \frac{1}{t} \sum_{k=1}^t (z_{k,i} - \hat{z}_{k,i})^2 \quad (33)$$

so that for each site i , we can fashion a column vector containing effective standard errors of each partial slope (34).

$$\text{SE}[\mathbf{b}_{t,i}] = \hat{\sigma}_{t,i} \cdot \sqrt{g_t} \cdot \sqrt{\text{diag}[\boldsymbol{\Pi}]} \quad (34)$$

One desiring a rigorous interpretation of effective standard errors would be advised to use the value of ρ that produces minimal prediction error, and also confirm the normality assumption of these errors.

5. “Local” and “Centered” Adaptive Least Squares

In cases where observations drift away from zero, the bias introduced by the regularizer in (23), may poorly serve to reduce prediction error. In such cases we can recursively center our

observations. The centered analogue of (21)-(24) is shown in (35)-(41).

$$g_t = (g_{t-1} + \rho) / (g_{t-1} + \rho + 1) \quad (35)$$

$$\widehat{\mathbf{B}}_t = \left(\mathbf{L}_{t-1}^\times - \bar{\mathbf{z}}_{t-1}^{*T} \bar{\mathbf{z}}_{t-1}^* + \mathbf{I}\lambda \right)^{-1} \left(\mathbf{L}_{t-1}^* - \bar{\mathbf{z}}_{t-1}^{*T} \bar{\mathbf{z}}_{t-1} \right) \quad (36)$$

$$\widehat{\mathbf{z}}_t = \bar{\mathbf{z}}_{t-1} + (\mathbf{z}^* - \bar{\mathbf{z}}_{t-1}^*) \widehat{\mathbf{B}}_t \quad (37)$$

$$\mathbf{L}_t^\times = \mathbf{L}_{t-1}^\times + g_t (\mathbf{z}^{*T} \mathbf{z}^* - \mathbf{L}_{t-1}^\times) \quad (38)$$

$$\mathbf{L}_t^* = \mathbf{L}_{t-1}^* + g_t (\mathbf{z}^{*T} \mathbf{z}_t - \mathbf{L}_{t-1}^*) \quad (39)$$

$$\bar{\mathbf{z}}_t^* = \bar{\mathbf{z}}_{t-1}^* + g_t (\mathbf{z}^* - \bar{\mathbf{z}}_{t-1}^*) \quad (40)$$

$$\bar{\mathbf{z}}_t = \bar{\mathbf{z}}_{t-1} + g_t (\mathbf{z}_t - \bar{\mathbf{z}}_{t-1}) \quad (41)$$

Notice firstly that we have moved our regularizer out of the update and into the assignment that creates our partial slopes, (36). Actually, when $\mathbf{L}_0^{(0)} = \mathbf{0}$, $\mathbf{L}_0^{(1)} = \mathbf{0}$, and $g_0 = \infty$, in the uncentered algorithm, we can do the same thing, that is move the regularizer from (23) to (22) without altering the predictions. Otherwise, the only distinction between the uncentered and centered solution is that in the latter we subtract out from the predictor, \mathbf{z}^* , and the predictand, \mathbf{z} , their respective estimated means at every recursion. The relationship, said in short, is best understood by recalling that in OLS, $\widehat{y}_i = b x_i = \bar{y} + b(x_i - \bar{x})$.

We lastly consider a third variant that we call ‘‘Local’’ ALS. Here, Local ALS entails running Uncentered ALS (21)-(24) n times — once for each site, but using *only* lagged responses for that site. For example, in a lag-2 forecast, the prediction $\widehat{z}_{t,i}$ is created with $\mathbf{z}^* = (z_{t-2,i}, z_{t-1,i})$ and $\mathbf{z}_t = z_{t,i}$. By implication, then, \mathbf{L}^\times would be 2×2 . In other words, we are viewing our data as n unique scalar time series, and, for some (ρ, λ) , we apply Uncentered ALS to each. Solving a system with Local ALS will be advantageous in cases of weak spacial correlation, or in cases where spacial correlation evolves too rapidly to be tracked by \mathbf{L}_t^\times and \mathbf{L}_t^* .

The method for producing the effective standard errors of a solution described at the end of the previous section can be readily used with either of these two variants.

6. Comparisons

While sometimes data arise through a known, well-defined system, data in nature usually do not. If we wish to forecast or predict, say, carbon monoxide concentrations over Boston from historical data, we do not choose a state-space model with the belief that we need only locate the correct hyperparameters to achieve minimal prediction error. Rather, we choose our model believing that it defines a rich enough class of relationships that truth and our fitted model will be close enough so that we may in some way profit from our effort. Empiric comparison between the performance of candidate solutions against ground truth offers a nice window by which we might view this profitability, not only because we can compare a typical cost between solutions, but because we can do so relative to the variation in cost associated with multiple realizations from the same system. Additionally, comparisons are often used as a forum by which different solutions may compete. Implicit in such presentations is that the constructed data in some way emulate something akin to a real dataset with which the audience may at some point come into contact. Certainly, the logic goes, if solution A is easier to implement than solution B, and provides similar quality predictions, the reader should turn to solution A in real life. This is, of course, contentious on its face, since there will always

be some lack of consensus over what fundamental assumptions are reasonable, e.g., what generative systems are worthy of consideration, and what candidate solutions are realistic.

With this said, our context here will be deliberately narrow and our advocacy of ALS somewhat conservative. First, we restrict ourselves to systems that are long in time, i.e., $\tau \gg n$. Second, the fitted Kalman solutions will be constructed in a way consistent with typical — though quite parsimonious — identification. Moreover, we present our results (boxplots and table) in a purely descriptive way; we do not seek to advance what is, ultimately, a qualitative interpretation by forwarding evidence inferentially over any solution’s true, say, mean RMSE (utilizing, e.g., t -tests).

6.1. Long in Time, Short in Space

We set $\boldsymbol{\omega}$ to be a row vector of $n = 11$ evenly spaced locations spanning $[0.3, 0.7]$. We use the R package **SSsimple**, Zes (2011), to simulate data for 12 hyperparameterizations of the system (1)-(2) — all combinations of

$$\mathbf{F} \in \{0.77\mathbf{I}, 0.999\mathbf{I}\}$$

$$\mathbf{R} \in \{1\mathbf{I}, 7\mathbf{I}\}$$

$$\mathbf{u} \in \{(2, 4, 6), (5, 10, 15, 20), (10, 20, 30, 40, 50, 60)\}$$

so that our observation-space function, \mathbf{H} , is a sine-cosine basis transformation using the frequencies contained in \mathbf{u} .

$$\mathbf{H}(\boldsymbol{\omega}, \mathbf{u}) = (\sin[\boldsymbol{\omega}^T \mathbf{u}], \cos[\boldsymbol{\omega}^T \mathbf{u}])$$

So, for example, when $\mathbf{u} = (2, 4, 6)$, \mathbf{H} will be $11 \times 3 \cdot 2 = 11 \times 6$.

For each we create 50 examples (replications) of length $\tau = 805$. We set our training range to $t \in 100:600$, and test over $t \in 601:800$. We ran three Kalman Solutions along with each of the three ALS variants, each ALS variant using lag-1, lag-2, and lag-3 information (detailed below). Figures 1-3 compare the resulting testing range RMSEs. **Please note** that the labels of the boxplots vary. What we have done is report only one boxplot for each of the three ALS variants — the one of the three lag solutions producing the mean lowest RMSE over the fifty forecasts. This was done to reduce clutter in the plots.

The forecasting method labelled “Oracle” refers to application of the Kalman solution using hyperparameters set to truth.

The forecasting method labeled “HBS” J “RF” indicates a solution where the observation matrix \mathbf{H} is estimated using J evenly spaced B-splines. We set $\mathbf{Q} = \mathbf{I}$, and assume that \mathbf{R} , \mathbf{F} are constant diagonal, with the constraints that $\text{Trace}[\mathbf{R}] > 0$, and that \mathbf{F} must be on or inside the unit circle, i.e., $\text{Trace}[\mathbf{F}] \leq J$. These two scalar values are found using Metropolis-Hastings (MH) to minimize the RMSE over the training range.

The forecasting method labeled “Hident RF” indicates a solution where $\mathbf{H} = \mathbf{I}$ and $\mathbf{Q} = \mathbf{I}$. We assume that \mathbf{R} and \mathbf{F} abide the constraints listed above. The two scalar hyperparameters were found using Metropolis-Hastings to minimize the RMSE over the training range. This solution ignores spacial association between responses: Setting $\mathbf{H} = \mathbf{I}$ ignores any deterministic covariation, and by setting \mathbf{R} to be diagonal, we ignore stochastic covariation.

The method “ALS L” k refers to the uncentered ALS algorithm shown in (21)-(24), with k being number of lagged observations used for our forecast. For example, “L2” means that $\mathbf{z}^* = (\mathbf{z}_{t-2}, \mathbf{z}_{t-1})$, and by implication, \mathbf{L}_t^\times is 22×22 .

The method “ALS vc L” k refers to the centered ALS variant given by (35)-(41).

The method “ALS loc L” k refers to the local ALS variant.

Video	Basis Frequencies	F	R	Z rng	Z var	O MSE	A MSE	O R^2	A R^2
Video	2,4,6	0.770	1	-11 to 8	8.1	4.05	4.10	49.65	49.12
Video	2,4,6	0.999	1	-56 to 2	351.7	4.15	4.27	98.82	98.79
Video	2,4,6	0.770	7	-15 to 11	14.1	10.47	10.67	25.58	24.17
Video	2,4,6	0.999	7	-59 to 3	357.8	10.95	11.46	96.94	96.80
Video	5,10,15,20	0.770	1	-12 to 11	10.8	5.22	5.31	51.47	50.65
Video	5,10,15,20	0.999	1	-59 to 51	1168.4	5.40	5.62	99.54	99.52
Video	5,10,15,20	0.770	7	-16 to 13	16.7	12.04	12.36	27.98	26.03
Video	5,10,15,20	0.999	7	-61 to 51	1174.5	12.93	13.68	98.90	98.84
Video	10,20,30,40,50,60	0.770	1	-11 to 12	15.5	7.43	7.50	52.10	51.63
Video	10,20,30,40,50,60	0.999	1	-88 to 111	2354.6	7.76	7.84	99.67	99.67
Video	10,20,30,40,50,60	0.770	7	-13 to 15	21.5	15.05	15.34	30.13	28.83
Video	10,20,30,40,50,60	0.999	7	-90 to 114	2360.6	16.82	17.16	99.29	99.27

Table 1: Quantitative results from our empiric comparisons. “Video” shows one of the system replications over the testing range. The ordinate range for each video can be found in Table column “Z rng.” White dots are the data, white path is true system mean function, gold path is Oracle *a priori* mean function estimate, red path is the mean function estimated through the B-spline basis. Column “Z var” gives the marginal MSE, averaged over all 50 replications, of our simulated responses over the testing range. Column “O MSE” gives the Oracle (Kalman *a priori*) MSE, averaged over all 50 replications, over the testing range. “A MSE” is the analogous MSE for the single **best** of all 9 ALS solutions. The last two columns, “O R^2 ” and “A R^2 ,” give the respective percentage of total variance explained, calculated in the usual way from the previous three columns. While it may be objectionable to consider the *marginal* percent of variance explained in stace-time data, it is nonetheless easy to conceptualize. In the two cases where the mean function is wiggly and loosely constrained (bottom row and thrid from bottom row), Centered ALS is the superior of the three ALS variants. Here the amount of total variance explained by ALS and the truthfully fitted KF (the Oracle) are numerically very close.

Early in our analyses we also employed the sub-optimal sub-space method to locate the full complement of hyperparameters, $\mathbf{H}, \mathbf{F}, \mathbf{R}, \mathbf{Q}$ as detailed in Ljung (1998) and implemented in the package **SSsimple**. These hyperparameters, when used in the Kalman solution occasionally produced decidedly inferior RMSEs, so these results were omitted.

In every case, whether fitting a Kalman solution or ALS, our hyperparameter space was deliberately chosen to be small — only 2 scalar, constrained values. This allowed for relatively fast fitting. Running MH on parallel threads, accomplished through the **R** package **snowfall**, we fit our Kalman solutions in roughly 25 seconds. It should be noted that in grey-box identification we need identify every element of $\mathbf{H}, \mathbf{F}, \mathbf{R}, \mathbf{Q}$, that is, $np + p^2 + n^2 + p^2$ (the number of elements in the respective matrices) total scalar values. Software implementations for this job are varied in their approach. In the case where $\mathbf{u} = (2, 4, 6)$, so $p = 6$, the **R** package **dIm**, Petris (2009), required several hours to locate our hyperparameters. The **R** package **Stem**, Cameletti (2009), can optimally fit a state-space model in reasonable time, but must be told \mathbf{H} in order to do so, and assumes an exponential covariogram structure in

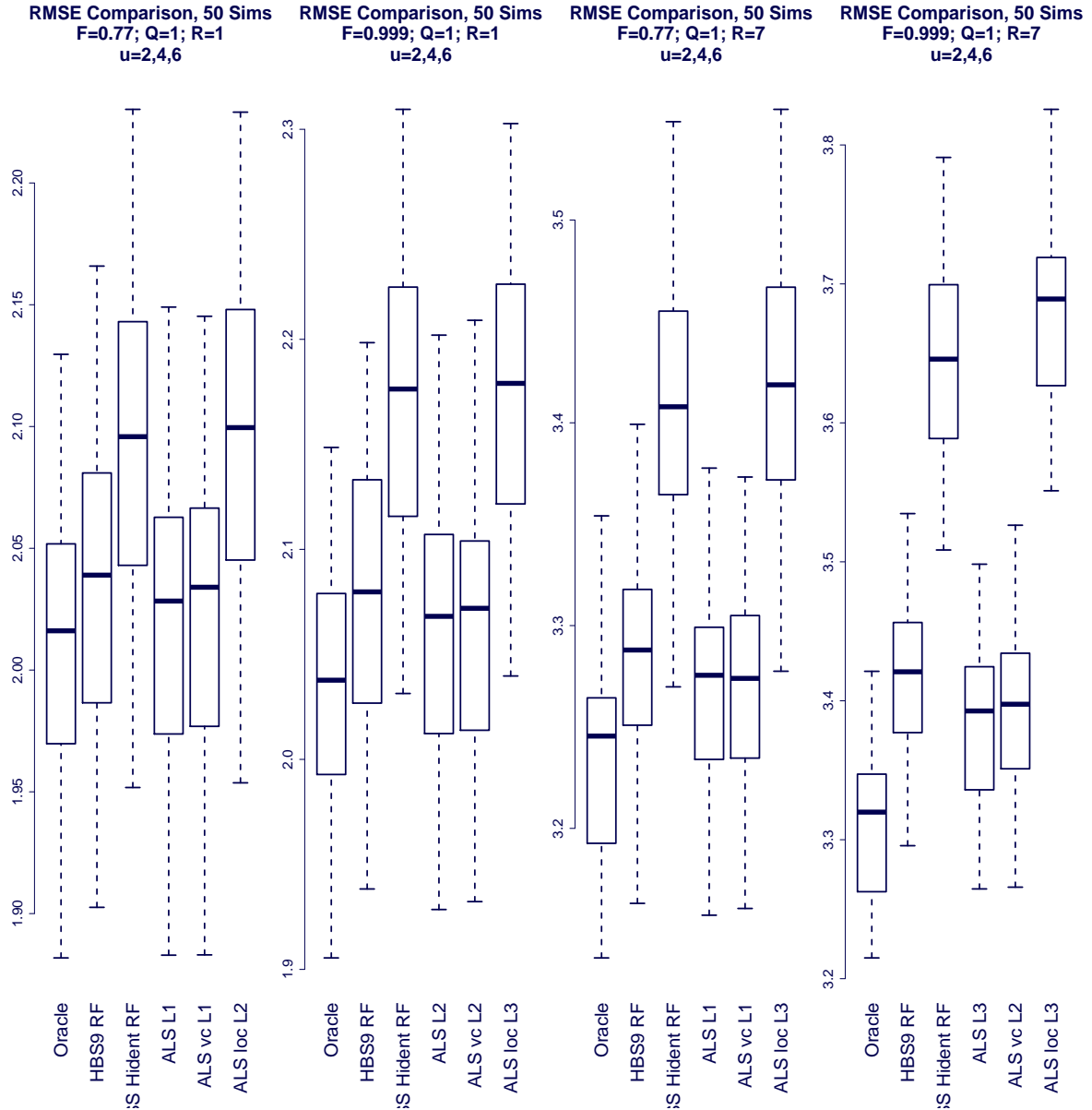


Figure 1: RMSE comparisons for generative systems with relatively smooth mean function. When the measurement error is low, $R = 1$, and the state is tightly constrained, $F = 0.77$, (left plot), the lag-1 Uncentered ALS forecast is of somewhat comparable quality to that of the Oracle. The greatest departure in performance between ALS and truth can be seen when the measurement noise is large, $R = 7$, and the state is allowed to wander, $F = 0.999$, (right plot). Here, if we think in terms of the total variability in RMSE between system replications, the 3rd quartile RMSE for the lag-3 Uncentered ALS is only a little better than the 1st quartile RMSE of the Oracle. We notice in every case at least one of the three ALS variants is superior to both of the parsimoniously fitted Kalman solutions.

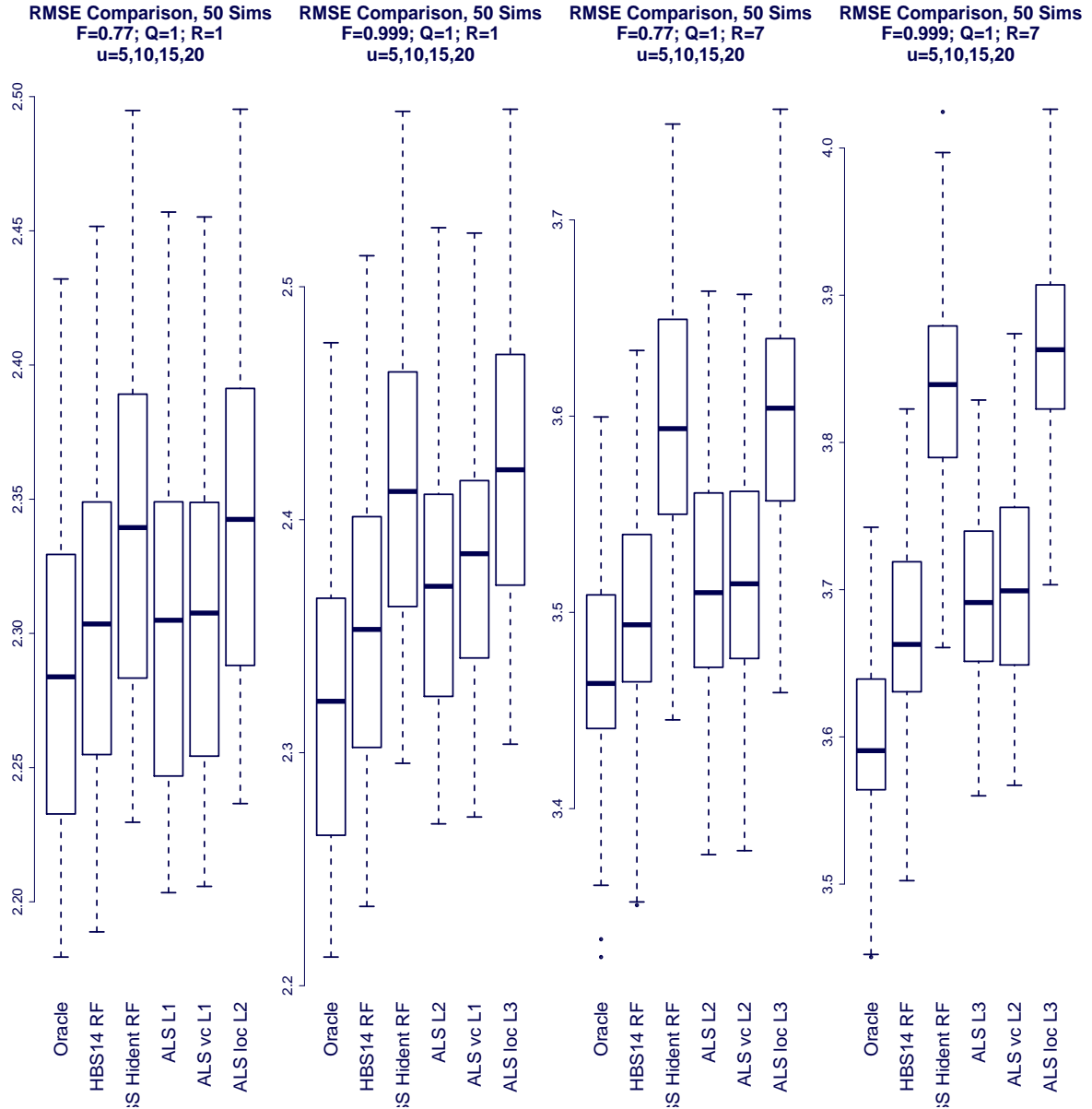


Figure 2: RMSE comparisons for generative systems with moderately wiggly mean function. Much like in Fig. 1, when the measurement error is low, $R = 1$, and the state is tightly constrained, $F = 0.77$, (left plot), the lag-1 Uncentered ALS forecast is somewhat comparable to that of the Oracle. And again, the greatest departure in performance between ALS and truth can be seen when the measurement noise is large, $R = 7$, and the state is allowed to wander, $F = 0.999$, (right plot). Here, if we think in terms of the total variability in RMSE between system replications, the 3rd quartile RMSE for the lag-3 Uncentered ALS is not quite as low as the 1st quartile RMSE of the Oracle. Notice in every case an ALS solution either beats or competes with both parsimoniously fitted Kalman solutions.

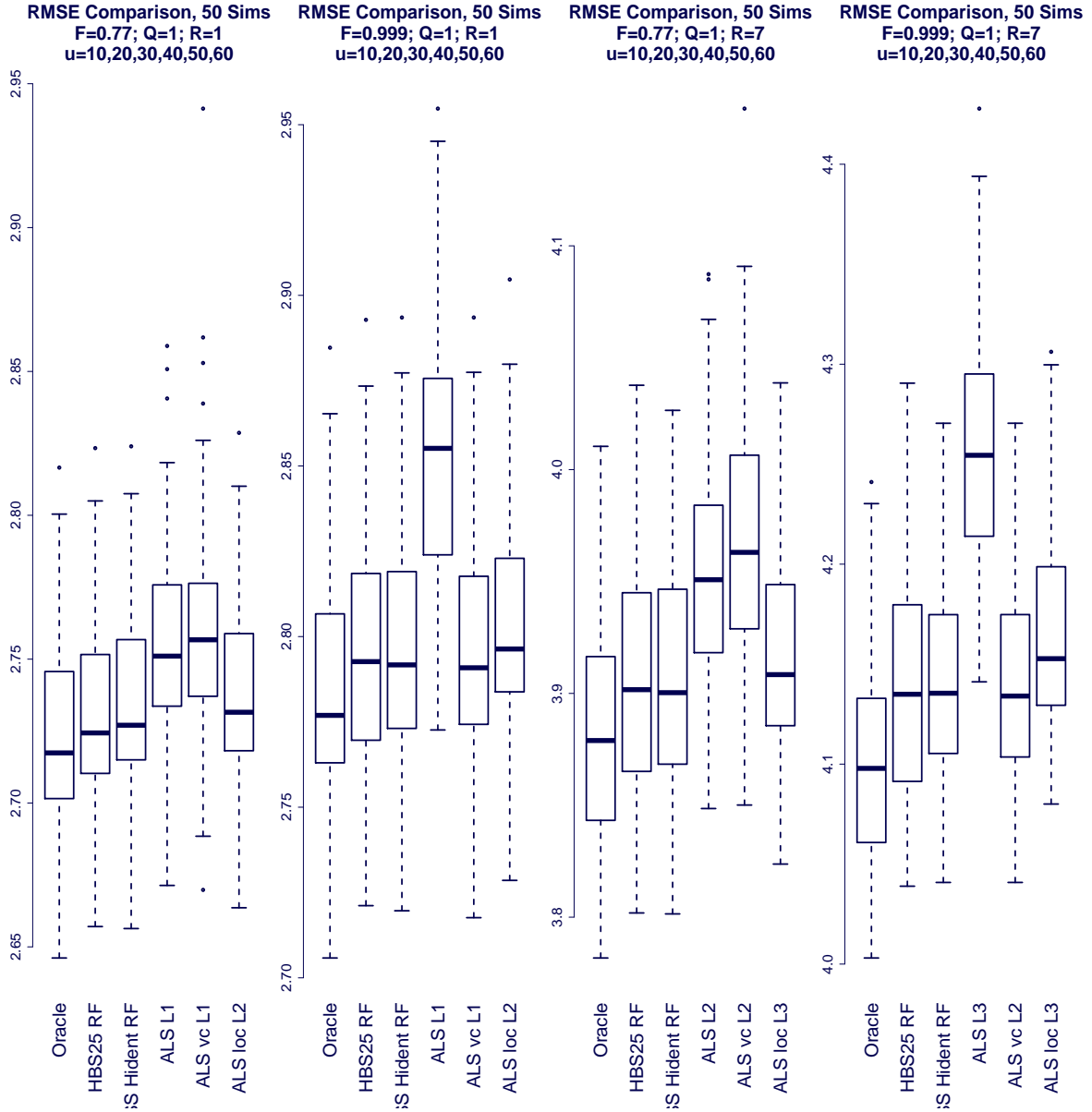


Figure 3: RMSE comparisons for generative systems with wiggly mean function. Notice that these results offer a different picture than in the previous cases. In both cases where the mean function is tightly constrained, $F = 0.77$, (first and third plots from right), Local ALS solutions bested the other two ALS variants, and provide forecasting RMSEs comparable with the parsimoniously fitted Kalman solutions. Under the other two generative systems, where the observations are allowed to wander from zero, $F = 0.999$, Centered ALS competes with the parsimoniously fitted Kalman solutions. This is spiritually consistent with the intuition offered in Section 5.

R. Additionally, the package **SSsimple** can be used in conjunction with the `optim()` function, but fitting time could be measured in terms of hours.

ALS, when using lag-2 information, and hence \mathbf{L}^\times is 22×22 , MH took about 15 seconds to locate ρ and the regularizer, λ . With lag-3 information, so \mathbf{L}^\times is 33×33 , MH took about 40 seconds to locate these two ALS hyperparameters.

6.2. The Error Surface

Being able to easily visualize a solution’s error surface is one luxury of having a small hyperparameter space. For both Centered on Uncentered ALS (running lag-2), we take the first realization from each of the twelve simulated systems and plot the testing range RMSE against (ρ, λ) . The results are offered as combined heat-contour maps collected together as TIFF booklets (links in text in the following paragraph). The domains for each hyperparameter are in log base 2, so that ρ runs from about one-millionth (2^{-20}) to about one-thousand, and the regularizer runs from about one-thousandth to about one-thousand. Whiteness in the heat map indicates relatively low RMSE, redness, relatively high RMSE. Some RMSE values appear in the contour bands.

Notice that while the **uncentered solution error surfaces** are not convex over our hyperparameter domain, the surface is nonetheless very well-behaved. In every case the testing RMSE is relatively uniformly low for small ρ and small λ (lower left corner of plot). This massive “sweet spot” suggests that the quality of ALS’s predictions will not deteriorate with modest departures in our two hyperparameters. A similar picture exists for the **centered solution error surfaces**, although here, in four cases (pages 2, 6, 10, 12), there exists large regions of possible false minima, i.e., bimodality of the error surface. When applying the lag-2 Centered ALS solution to these systems, one may fit ALS with large ρ and λ and obtain a low RMSE. This is the classic “double-edged sword,” since while it may be possible that MH gets caught in a false minima, ALS is robust enough so that it may still perform quite well even with erroneous hyperparameters.

7. Applications

Here we apply ALS, as well as the Kalman Filter, to three datasets, with some attention to methods of visualization.

“Eclipse” videos show value of forecasted and actual values as diameter of orange over red circles respectively. The amount of red one sees, and whether the residual red is inside or outside the orange circles offers a crude visual estimate of magnitude and direction of forecast error.

The “Correlation Star” videos show, for each site in turn over some time range, the estimated ALS site-wise correlation between that site and its $n-1$ cohort sites, with magnitude indicated by color. For some value of (ρ, λ) , \mathbf{L}_t^\times is extracted (after the update) from the ALS algorithm (21)-(24). If, for example, we use lag-1 information to forecast, i.e., $\mathbf{z}^* = (\mathbf{z}_{t-1})$, then \mathbf{L}_t^\times will be the *a priori* ALS estimate of $\text{Cov}[\mathbf{z}_t^T, \mathbf{z}_t]$ at time t . We call this the lag 0 \times lag 0 covariance estimate. This estimated covariance matrix is converted into a correlation matrix. The i th column of this correlation matrix contains the ALS estimated spacial correlations between site i and its cohort sites.

“Maximal Correlation” videos show, for each site simultaneously over time, the site with which it maximally spatially correlates, according to ALS, by connecting the two sites with a line segment. Magnitude is represented as line thickness. The computational mechanics are identical to that of the correlation star.

The purpose of the two correlation plots is potentially threefold. Firstly, as a means of data visualization, secondly, as a diagnostic tool, and thirdly, as precursors to covariogram selection for the sake of spacial interpolation. It should be stressed that, like many visual diagnostic tools in common use by the statistician to aid inference, these are quite subject to qualitative implementation and interpretation. Consider, for example, that since these plots are created through the observation covariances, \mathbf{L}^\times , which are themselves estimated over time, the researcher, in their creation, has free choice of choosing the update weighting hyperparameter, ρ , to serve the desired visualization. One natural choice (as we’ve done here) is to use the ρ that minimizes the RMSE. However, by increasing ρ , the researcher can visually emphasize the spacial correlation of more recent past observations.

7.1. California Ozone

We examine atmospheric Ozone (O3) over California: Daily observations of the average one-hour maximum concentration in ppb for 47 sites from January 1st, 2007, to December 31st, 2009. Data are provided by the California Air Resources Board (CARB). The grand standard deviation is 17.76, median is 45, minimum is 2, maximum is 176, units ppb.

We set our training region to be $t \in 100:730$, and our testing region to be $t \in 731:1096$.

If we use the mean of \mathbf{Z} over the testing region as a forecast for all observations in the test range, we get an RMSE of 17.79. If we use $z_{t-1,i}$ to predict $z_{t,i}$ for all values t in the testing range, we get an RMSE of 10.88.

ALS

We forecast one step ahead using 2-lag observations using (21)-(24). We used MH over our training range to locate $(\rho, \lambda) = (5.968 \cdot 10^{-6}, 60.00)$ in a little under 8 minutes. Testing with these hyperparameters produces an RMSE of 9.438.

We readily notice the annual seasonal trend in the eclipse plot; temperature is a well understood precursor for O3 formation. We can slightly improve our 1-step-ahead forecast by including a sine-cosine term in the system covariances, \mathbf{L}^\times and \mathbf{L}^* . This is achieved by having $\mathbf{z}^* = (\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, a \cdot \sin[2\pi t/365.25], a \cdot \cos[2\pi t/365.25])$, with a set to 21. Adding this seasonal information slightly dropped our testing RMSE to 9.383, with $(\rho, \lambda) = (5.424 \cdot 10^{-6}, 57.08)$ — (training takes about 9 minutes). Analogously applying Centered ALS dropped the testing range RMSE to 9.294, with $(\rho, \lambda) = (5.511 \cdot 10^{-6}, 51.74)$ — (training takes about 10 minutes).

The error surfaces for both the centered and uncentered seasonal ALS solutions reveal a large sweet spot around our located hyperparameters.

Kalman Solution

Since we are forecasting to monitored locations, we start by creating the $n \times n$ $\mathbf{H}_1 = \mathbf{I}$. The statistician, keen on viewing the measurement function, \mathbf{H} , as a design matrix, will notice directly that \mathbf{H}_1 is a “dummy” variable that uniquely indicates each of the n sites. We create \mathbf{H}_t as a time-varying measurement function by appending two columns to \mathbf{H}_1 :

$\mathbf{H}_t = (\mathbf{H}_1, \mathbf{s}(t), \mathbf{c}(t))$, with $\mathbf{s}(t)$ being a constant column vector of height n with each element equalling $\sin[2\pi t/365.25]$, and $\mathbf{c}(t)$ analogously constructed using the cosine function.

We parsimoniously assume $\mathbf{Q} = \mathbf{I}$, $\mathbf{R} = \text{diag}[\mathbf{R}]$, $\mathbf{F} = \text{diag}[\mathbf{F}]$.

We use MH on `SS.solve.tv()` in the R package `SSsimple` to locate our two hyperparameters over the training region in about 1 minute to find $(\mathbf{R}, \mathbf{F}) = (3.456, 0.9857)$, producing a testing RMSE of 10.125.

* Daily California Ozone *

Centered ALS 1-step-ahead forecast “Eclipse” video using lag-2 information along with the seasonal sine-cosine terms. For these predictions (unlike in the comparisons above), we use hyperparameters located by training over almost the full data, $t \in 100 : 1093$: Magnitude of actual values are shown as size of red circles, magnitude of forecasted values shown as size of orange circles. Notice, for example, the day August 11, 2009. Many sites around the San Francisco Bay were overestimated (red shows inside the orange), while a few values near and to the east of Los Angeles were underestimated (some red shows outside the orange circle). Also notice, for another example, on July 7, 2008. ALS profoundly underestimated ozone concentrations near San Francisco Bay.

* O3 Correlation Star *

Correlation star video, lag-0 \times lag-0. For example, scroll forward to time 1:07, then find the date August 9th of 2009. Here we have correlation strengths (by color indicated in the legend) between a location in San Francisco and the other 26 cohort sites. Notice firstly, that all correlations appear strongly positive. However, we find the suggestion of anisotropy: Consider the fan of lines indicating estimated correlation to sites in southern California. Here, correlation for roughly equidistant sites run from about 0.87 (bluish lines running to inland sites), to roughly 0.97 (orange lines running to coastal sites).

* O3 Maximal Correlation *

Maximal correlation video, lag-0 \times lag-0. We first recall some California geography: The Sierra Nevada mountain range roughly bisects the state along its length. With this in mind a fairly clear pattern comes into view. Coastal sites tend to maximally correlate with other coastal sites, and mountainous sites tend to maximally correlate with other mountainous sites.

7.2. Irish Gale

Here we examine wind speed over Ireland: Daily average wind speed in meters-per-second for 11 sites from January 1st, 1961, to December 31st, 1978, for a total of $\tau = 6574$ days. The grand standard deviation of \mathbf{Z} is 2.900, median is 4.779, minimum is 0, maximum is 21.88, units meters-per-second.

We set our training region to be $t \in 100:4000$, and our testing region to be $t \in 4001:6571$.

If we use the mean of \mathbf{Z} over the testing region as a forecast for all observations in the test range, we get an RMSE of 2.921. If we use $z_{t-1,i}$ to predict $z_{t,i}$ for all values t in the testing range, we get an RMSE of 2.408.

ALS

We forecast one step ahead using 2-lag observations. Using MH over our training range we locate $(\rho, \lambda) = (1.384 \cdot 10^{-6}, 0.1908)$ in about 44 seconds. Testing produced an RMSE of 2.094.

As with ozone, wind is seasonal. We can very slightly improve our 1-step-ahead forecast by including a sine-cosine term in the system covariances, \mathbf{L}^\times and \mathbf{L}^* . This is achieved by having $\mathbf{z}^* = (\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, a \cdot \sin[2\pi t/365.25], a \cdot \cos[2\pi t/365.25])$, with a set to 3. Analogous training and testing produced an RMSE over the testing range of 2.088, with $(\rho, \lambda) = (1.268 \cdot 10^{-6}, 0.2080)$ — (training takes about 50 seconds). Application of Centered ALS dropped the testing range RMSE to 2.033, with $(\rho, \lambda) = (9.370 \cdot 10^{-7}, 0.2736)$ — (training takes about 1 minute).

Just as with California Ozone, the **the error surfaces** for both the centered and uncentered seasonal ALS solutions reveal a large sweet spot around our located hyperparameters.

Kalman Solution

We proceed in a manner identical to that described above for California Ozone. We find that $(\mathbf{R}, \mathbf{F}) = (10.90, 0.9739)$ gave a testing RMSE of 2.2282 — (training takes about 2 minutes).

*** Daily Irish Wind Speed ***

Centered ALS 1-step-ahead forecast “Eclipse” video using lag-2 information along with the seasonal sine-cosine terms. One rather evident trait of our forecasts is that, unlike with California Ozone, daily under- or over-estimation is by and large independent of location. That is, if on any one day we have over-estimated our forecast for one site, we have probably done so for all sites. Notice the date November 11, 1973. Here our forecast rather dramatically under-estimated the Irish gale.

*** Irish Wind Speed Correlation Star ***

Correlation star video, lag-0 \times lag-0. Perhaps the holistic message here is the general isotropy of daily average wind speed over Ireland — at least as far as the year 1978 goes. By-and-large, the strength of correlation appears dependent on distance, and little on direction.

*** Irish Wind Speed Maximal Correlation ***

Maximal correlation video, lag-0 \times lag-0. Here, we find additional visual evidence for isotropy. Sites tend to maximally correlate with the nearest, or close to nearest site.

7.3. Global Methane

We obtained atmospheric methane (CH₄) concentrations in units ppb from [World Data Centre for Greenhouse Gases](#) for 23 worldwide sites recorded monthly from January 1983 to December 2007, for a total of $\tau = 300$ epochs. The raw data, however contained 478 missing observations. We imputed missing values using a smoothing ALS variant, readily constructed by having $\mathbf{z}^* = (\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1})$. This implementation requires one extra step. Notice that we’re using \mathbf{z}_t to predict itself. In order to avoid having our ALS algorithm merely duplicate existing values,

at each time t , we run the ALS solution n times, once for each site, each time removing the respective row and columns from our predictor and covariance estimates (42).

$$\hat{z}_{t,i} = \mathbf{z}_{[-i]}^* \left(\mathbf{L}_{[-i,-i]}^\times \right)^{-1} \mathbf{L}_{[-i,i]}^* \quad (42)$$

We initially fill missing values with a global mean value, then use MH over $t \in 50:295$ to simultaneously impute and search out hyperparameters.

On the completed dataset we set our training region to be $t \in 50:250$, and our testing region to be $t \in 251:295$.

If we use the mean of \mathbf{Z} over the testing region as a forecast for all observations in the test range, we get an RMSE of 53.29. If we use $z_{t-1,i}$ to predict $z_{t,i}$ for all values t in the test range, we get an RMSE of 10.27.

ALS

Forecasting one step ahead using 2-lag observations: Training locates $(\rho, \lambda) = (0.00516, 30.89)$ in about 10 seconds. Testing produced an RMSE of 6.538.

Again, we have a seasonal effect. We can improve our 1-step-ahead forecast by including a sine-cosine term in the system covariances, \mathbf{L}^\times and \mathbf{L}^* . This is achieved by having $\mathbf{z}^* = (\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, a \cdot \sin[2\pi t/12], a \cdot \cos[2\pi t/12])$, with a set to 100. Analogous training (about 11 seconds) and testing produced an RMSE over the testing range of 6.376, with $(\rho, \lambda) = (0.01354, 40.73)$.

Running Centered ALS produced a comparable RMSE.

The error surfaces for our centered and uncentered ALS solutions are decidedly appealing. It appears that almost any straight line passing through the minima maps to a convex error curve.

Kalman Solution

We proceed in a manner identical to that described above for California Ozone. We find that $(\mathbf{R}, \mathbf{F}) = (1 \cdot 10^{-7}, 1)$ gave a testing RMSE of 9.864 — (training takes about 13 seconds).

* Global Methane *

ALS 1-step-ahead forecast CH4 concentrations in ppb on globe, using lag-2 and seasonal sine-cosine information. This video showcases one possible means of expressing location-wise magnitude on the globe — as opposed to, say, a 2D projection map. The colored disks are centered at the monitored site, color indicates magnitude as shown in the legend. The forecasts clearly track the upward trend of atmospheric methane over the *circa* 22 year period.

8. Final Thoughts

The gut appeal of the two-level hierarchical model we've presently considered, (1)-(2), should be owed to its possession of key ingredients for describing action in space-time. The measurement function, \mathbf{H} , defines a transform by which the latent state maps mean observation values over space. The state noise variance, \mathbf{Q} , determines the amount of “temporal flexibility” of

this mean function, while the measurement noise variance, \mathbf{R} , local covariation, and \mathbf{F} , the sensitivity to long-term drift in the magnitude of the observations. In our present work, we focused exclusively on forecasting to monitored locations, and so having $\mathbf{H} = \mathbf{I}$, or having \mathbf{H}_t contain \mathbf{I} , is usually a sensible place to start fitting data. In practice, though, whether for the sake of interpolating, or simply imposing spatial structure, two-dimensional space can make selection of a suitable \mathbf{H} challenging. We can impose an additive assumption and have $\mathbf{H}(\boldsymbol{\omega}_{xy}) = [\mathbf{H}(\boldsymbol{\omega}_x), \mathbf{H}(\boldsymbol{\omega}_y)]$, but this simplification often yields dissatisfying mean functions. Nonetheless, in application the statistician may call upon the vast literature of non-parametric function estimation, e.g., Wasserman (2006), Efromovich (1999), Abramovich, Bailey, and Sapatinas (2000), to construct a suitable measurement function.

One appealing feature of ALS is its tiny hyperparameter space. Not only does this allow for relatively fast fitting, it also suggests the prospect of embedding of ALS in larger routines. This fact, coupled with its flexibility should fire the researcher’s imagination. Large systems can be decomposed into smaller ALS solutions run in parallel or in series. In fact, in concept, at least, outputs and inputs may be tied between two or more ALS solutions — a process potentially comprising a hierarchy of latent and exogenous constructs similar to what one sees in some structural equation models (SEMs) — see Pearl (2010) or Pearl (2009) for background on SEMs. The small hyperparameter space means one may fit ALS with a stochastic method such as Metropolis-Hastings (as we’ve done here). This in turn means that we may fit with an arbitrary cost function. In light of the growing pragmatics of ecology, for example, one may wish to formulate an objective function with the costs and benefits associated with regional attainment. We also saw that ALS, by its very construction, provides objects that are both instructive and diagnostic. By visualizing the updated covariances, \mathbf{L}_t^\times and \mathbf{L}_t^* , the researcher may acquire insight into how spatial covariation evolves in time.

References

- Abramovich F, Bailey TC, Sapatinas T (2000). “Wavelet Analysis and Its Statistical Applications.” *The Statistician*, **49**, 1–29.
- Cameletti M (2009). *Stem: Spatio-temporal models in R*. R package version 1.0.
- Efromovich S (1999). *Nonparametric Curve Estimation (Springer Series in Statistics)*, volume 1862. Springer U.S., New York.
- Gneiting T, Genton MG, Guttorp P (2005). “Geostatistical Space-Time Models, Stationarity, Separability and Full Symmetry.” *Technical report*, University of Washington.
- Gunnarsson S (1994). “On Covariance Modification and Regularization in Recursive Least Squares Identification.” *In 10th IFAC Symposium on System Identification — SYSID’94*, **2**, 661–666. Copenhagen, Denmark.
- Gunnarsson S (1996). “Combining Tracking and Regularization in Recursive Least Squares Identification.” *Proceedings of the 35th Conference on Decision and Control*. Kobe, Japan.
- Haykin S (2002). *Adaptive Filter Theory*. Forth edition. Prentice-Hall, Upper Saddle River.

- Kyriakidis PC, Journel AG (1999). “Geostatistical Space-Time Models: A Review.” *Mathematical Geology*, **31**(6), 651–684.
- Leung CS, Young G, Sum J, Kan W (1999). “On the Regularization of Forgetting Recursive Least Square.” *IEEE Transactions on Neural Networks*, **10**(6), 1482–1486.
- Ljung L (1998). *System Identification: Theory for the User (2nd Edition)*. Prentice Hall PTR.
- McCulloch JH (2005). “The Kalman Foundations of Adaptive Least Squares, With Application to U.S. Inflation.” *Unpublished*.
- Pearl J (2009). *Causality: Models, Reasoning, and Inference*. 2nd edition. Cambridge University Press, New York.
- Pearl J (2010). “The Causal Foundations of Structural Equation Modeling.” *Technical Report, R-370, August*. URL http://ftp.cs.ucla.edu/pub/stat_ser/r370.pdf.
- Petris G (2009). *d1m: Bayesian and Likelihood Analysis of Dynamic Linear Models*. R package version 1.0-2, URL <http://CRAN.R-project.org/package=d1m>.
- Sayed AH (2003). *Fundamentals of Adaptive Filtering*. John Wiley & Sons, Hoboken, N.J.
- Shumway RH, Stoffer DS (2006). *Time Series Analysis and Its Applications, With R Examples*. Second edition. Springer, N.Y.
- Stein ML (2002). “The Screening Effect in Kriging.” *The Annals of Statistics*, **30**(1), 298–323.
- Wasserman L (2006). *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- West M, Harrison J (1997). *Bayesian Forecasting and Dynamic Models*. Second edition. Springer-Verlag, New York.
- World Data Centre for Greenhouse Gases (—). URL <http://gaw.kishou.go.jp/wdcgg/>.
- Zes D (2011). *SSsimple: Solve, Fit, Simulate State Space Systems*. R package version 0.5, URL <http://CRAN.R-project.org/package=SSsimple>.

Affiliation:

Dave Zes
Department of Statistics, UCLA
Los Angeles, CA 90095-1554
E-mail: davezes@stat.ucla.edu
URL: <http://www.stat.ucla.edu/~davezes>

Journal of Environmental Statistics
Volume VV, Issue II
MMMMMM YYYY

<http://www.jenvstat.org>
Submitted: yyyy-mm-dd
Accepted: yyyy-mm-dd
