Marginally Parametrized Spatio-Temporal Models and Stepwise Maximum Likelihood Estimation

Matthew Edwards,* Stefano Castruccio† and Dorit Hammerling‡

July 2, 2018

Abstract

In order to learn the complex features of large spatio-temporal data, models with large parameter sets are often required. However, estimating a large number of parameters is often infeasible due to the computational and memory costs of maximum likelihood estimation (MLE). We introduce the class of marginally parametrized (MP) models, where inference can be performed efficiently with a sequence of marginal (estimated) likelihood functions via stepwise maximum likelihood estimation (SMLE). We provide the conditions under which the stepwise estimators are consistent, and we prove that this class of models includes the diagonal vector autoregressive moving average model. We demonstrate that the parameters of this model can be obtained at least three orders of magnitude faster using SMLE compared to MLE, with only a small loss in statistical efficiency. We apply a MP model to a spatio-temporal global climate data set (in order to learn complex features of interest to climate scientists) consisting of over five million data points, and we demonstrate how estimation can be performed in less than an hour on a laptop.

Keywords: spatio-temporal model, massive data, stepwise estimation, parallel computation.

*School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom (E-mail: m.edwards3@ncl.ac.uk).
†Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, United States of America (E-mail: scastruc@nd.edu).
‡Department of Analytics and Integrative Machine Learning, National Center for Atmospheric Research, Boulder, CO 80307, United States of America (E-mail: dorith@ucar.edu).
1 Introduction

The ever-increasing availability of large and complex spatio-temporal data in many areas of science including census, global climate, neuroscience and epidemiology, calls for the development of flexible models able to capture the features of interest to stakeholders. A model flexible enough to learn structural dependence at multiple scales, however, comes at the cost of large parameter sets required to describe it, and hence makes traditional inferential approaches such as maximum likelihood estimation (MLE) often infeasible. MLE depends on both the computational and memory costs of evaluating the likelihood function and the number of evaluations of the optimization algorithm, and while the costs of each evaluation increase with the number of data points, the number of evaluations increases with the number of parameters. It is therefore essential to consider the costs of evaluation and the number of evaluations when specifying models with large parameter sets for large complex spatio-temporal data.

For a Gaussian model, if the data set consists of $N$ data points and the covariance matrix is unstructured, then these costs are $O(N^3)$ floating point operations (flops) and $O(N^2)$ memory (in the non-Gaussian case, the cost is typically considerably higher: a max-stable likelihood evaluation requires $O(B_N)$ flops, where $B_N$ is the Bell number of order $N$, see Castruccio et al. (2016)). Approaches that reduce these costs often use structured covariance matrices (e.g sparse, low-rank, circulant) (Golub and Van Loan 2012, Davis 1979) to leverage numerical linear algebra routines (e.g. FFT) and parallel computation. For large spatio-temporal data, approaches include: weighted composite likelihoods (Bevilacqua et al., 2012), full-scale approximations (Zhang et al., 2015), dynamic nearest neighbour processes (Datta, Banerjee, Finley, Hamm, Schaap et al., 2016) and dynamic multi-resolution spatial models (Johannesson et al., 2007). In certain cases, approaches for large spatial data can be applied to the innovations of dynamic models (e.g. VARMA models). These approaches include: fixed rank kriging (Cressie and Johannesson, 2008), lattice kriging (Nychka et al., 2015), predictive processes (Banerjee et al., 2008), covariance tapering (Kaufman et al., 2008), multi-resolution approximations (Katzfuss, 2017), nearest neighbour processes (Datta, Banerjee, Finley and Gelfand, 2016) and stochastic partial differential equations (Lindgren et al., 2011).
All these approaches focus on models (possibly misspecified, in the case of composite likelihoods) with likelihood functions that have low costs of evaluation, hence their applicability to large spatio-temporal data. While these methods have been proven to be scalable for a wide range of applications, the ability to learn complex spatio-temporal patterns could come at the price of large parameter sets, and hence it is essential to consider the number of evaluations as well as the costs of evaluation. The number of evaluations depends on the number of iterations of the optimization algorithm (e.g. two evaluations per iteration for a numerical derivative) which, in turn, depends on the number of parameters \( P \) (Vavasis, 1991). If the likelihood function is concave then algorithms exist where the number of iterations is polynomial in \( P \) (e.g. ellipsoid algorithm, Yudin and Nemirovski˘ı, 1977). However, if the likelihood function is non-concave, algorithms only exist where the number of iterations is exponential in \( P \) (Yudin and Nemirovski˘ı, 1983). Therefore, MLE is infeasible for models with large parameter sets if the likelihood function is non-concave. In order to estimate a large number of parameters, parameter subsets must be estimated in multiple steps via a stepwise estimation method.

Multi-stage approaches are stepwise estimation methods often used in applications with different types of dependence (e.g. temporal, spatial) such as non-linear mixed models (Giltinan and Davidian, 1995). Schabenberger and Gotway (2017, pp. 431-433) proposed a two-stage approach where time series submodels are fit at each spatial location and then a spatial model is fit to the combined residuals. A large class of multi-stage approaches has been proposed in the context of global climate data (Castruccio and Stein, 2013; Castruccio and Guinness, 2017) where estimation in time, longitude and latitude is performed in three stages. An extension has been proposed in Castruccio and Genton (2016) where estimation of altitude dependence is performed in a fourth stage. In the context of neuroscience a multi-stage approach for whole-brain data was proposed in Castruccio et al. (2018). These classes of models have a reduced evaluation cost, as each submodel is defined over fewer data points. They also require a reduced number of evaluations, as each submodel depends on fewer parameters. Therefore, they are ideal for dealing with large and complex spatio-temporal data.

These multi-stage methodologies have emerged in the aforementioned statistical liter-
ature as a collection of *ad-hoc* methods for large complex spatio-temporal data where the nature of the problem suggests that modeling at multiple resolutions is useful for learning properties of interest. The lack of an underlying and unifying framework has so far prevented a full understanding of the generalizability of such models. Perhaps more importantly, a multi-stage approach does not guarantee an underlying joint model, and hence does not always allow probabilistic statements about the data. Without a joint model, minimum mean squared error prediction cannot be performed as conditional models cannot be derived. This is particularly problematic in spatio-temporal statistics where prediction is often the primary objective, but it is also a concern in simulation as without a joint model there is no explicit distribution to sample. Furthermore, the lack of a proper class of models did not allow for providing any asymptotic consistency results.

The aim of this work is to provide the foundation for multi-stage approaches by defining a joint model that can then be applied to large complex spatio-temporal data. This very general class of models is termed marginally parametrized (MP), does not require any Gaussian or stationary assumption, and is such that multi-stage inference is always possible with submodels that are proper marginal distributions. We define the general stepwise estimation method, termed stepwise maximum likelihood estimation (SMLE), and we provide the conditions that a MP model must satisfy to achieve asymptotic consistency across all stages of the inference.

The remainder of this paper is structured as follows: in Section 2 we define the class of MP models and a spatio-temporal model that is a member of this class, i.e. the diagonal VARMA model. In Section 3, we introduce the SMLE method, the asymptotic consistency theorem and the application of SMLE to the diagonal vector autoregressive moving average (VARMA) model. In Section 4 we provide two simulation studies. The first compares MLE with SMLE for the diagonal VARMA model with Matérn covariance \( \text{Stein}, 1999 \) and the second corroborates the consistency theorem for the corresponding estimators. In Section 5 we apply the diagonal VARMA model to a large complex spatio-temporal global climate data set. We conclude in Section 6.
2 Marginally Parametrized Model

Here we introduce the class of marginally parametrized (MP) models and a spatio-temporal model that is a member of this class, i.e. the diagonal VARMA model.

2.1 Definition and Heuristics

Denote $y$ a data set consisting of $N$ data points and denote $\mathcal{L}(y \mid \theta)$ a corresponding joint likelihood function that depends on a parameter vector $\theta$ consisting of $P$ parameters.

**Definition 1** (Marginally Parametrized Models). A model for $y$ is MP if there exists a finite sequence of $K > 1$ data subsets $(y_k)$ such that the marginal model of $y_k$ depends on a parameter subset with a partition $\theta_k, \eta_k$ where $\theta_k \neq \emptyset$ and $\eta_k \subseteq \theta_1 \cup \cdots \cup \theta_{k-1}$ ($\eta_1 = \emptyset$) for $k = 1, \ldots, K$ and $\theta_1, \ldots, \theta_K$ is a partition of $\theta$.

The sequence of data subsets $(y_k)$ corresponds to a sequence of marginal models. Each marginal model depends on a parameter subset that is partitioned into a set of primary and nuisance parameters ($\theta_k$ and $\eta_k$ respectively). The primary parameters correspond to the parameters previous marginal models in the sequence do not depend on; whereas, the nuisance parameters correspond to the parameters previous models in the sequence do depend on. Heuristically, the primary parameters of each marginal model control the dependence only within data subsets; hence, marginally parameterized. As a consequence of these conditions, the parameter set of a MP model can be estimated with a sequence of marginal (estimated) likelihood functions (Pawitan, 2001, Section 10.6), see Subsection 3.1.

2.2 Diagonal VARMA Model

We now consider a special case where $y$ is a spatio-temporal data set where the sampling design consists of $T$ regularly spaced time points $t \in \mathbb{Z}_{>0}$ at $S$ arbitrarily spaced, but fixed in time, locations $x_s \in \mathcal{M}^d$, where $\mathcal{M}^d$ is a $d$-dimensional manifold (e.g. plane, sphere); for a total of $N = S \cdot T$ data points. Denote $Y_t$ the random vector corresponding to the data points at all the locations at time point $t$. The diagonal VARMA model (Lütkepohl, 2005)
with autoregressive (AR) order $p$ and moving average (MA) order $q$ is defined as

$$Y_t = \mu + \sum_{i=1}^{p} \Phi_i Y_{t-i} + \sum_{j=1}^{q} \Pi_j \Sigma U_{t-j} \quad \text{where} \quad U_t \sim \mathcal{N}(0, R(\nu)), \quad (1)$$

where $\mu = (\mu_1, \ldots, \mu_S)$ is the vector of mean parameters, $\Sigma = \text{diag}(\sigma_s)$ is the diagonal matrix of standard deviation parameters, $\Phi_i = \text{diag}(\phi_{i,s})$ are the diagonal matrices of AR parameters and $\Pi_j = \text{diag}(\pi_{j,s})$ are the diagonal matrices of MA parameters. Furthermore, $U_t$ are i.i.d. in time, centered and unscaled Gaussian innovations and $\nu$ is the set of correlation parameters. This is a very flexible model: the mean, standard deviation, AR and MA parameters can be different at each location and the correlation matrix $R$ has no constraints.

To prove that (1) is a MP model we must demonstrate that there exists a finite sequence of data subsets that satisfy the conditions of Definition 1. Consider the finite sequence of $K = S + 1$ data subsets $(y_k)$ where $y_k$ is the time series at spatial location $x_k$ for $k$ from 1 to $S$ and $y_K = y$. The marginal model of $y_k$ is an ARMA model that depends on a parameter subset with a partition $\theta_k, \eta_k$ where

$$\theta_k = \mu_k \cup \sigma_k \cup \left( \bigcup_{i=1}^{p} \phi_{i,k} \right) \cup \left( \bigcup_{j=1}^{q} \pi_{j,k} \right)$$

and $\eta_k = \emptyset$ for $k$ from 1 to $S$. The proof is provided in Appendix A. Note that this model is MP as a result of the AR and MA parameter matrices being diagonal. The marginal (joint) model of $y_K$ is a diagonal VARMA model that depends on a parameter subset with a partition $\theta_K, \eta_K$ where $\theta_K = \nu$ and $\eta_K = \theta_1 \cup \cdots \cup \theta_S$. Clearly, this finite sequence of data subsets satisfies the conditions of Definition 1.

## 3 Inference

This section introduces the SMLE method for estimating the parameter set of a MP model in multiple steps, the SMLE consistency theorem and the details of SMLE for the diagonal VARMA model.
3.1 Stepwise Maximum Likelihood Estimation

Assume that $L(\theta \mid y)$ is a MP model likelihood function for $y$ and denote $(y_k)$ the corresponding finite sequence of data subsets that satisfies Definition 1. Let $L_k(\theta_k, \eta_k \mid y_k)$ denote the marginal likelihood function of $y_k$, that depends on a parameter subset with partition $\theta_k, \eta_k$ for $k = 1, \ldots, K$. Instead of estimating $\theta$ with the MP joint likelihood function $L(\theta \mid y)$ in one step (MLE), the SMLE method estimates $\theta_1$ with the marginal likelihood function $L(\theta_1 \mid y)$ in step one (since $\eta_1 = \emptyset$) and estimates $\theta_k$ with the marginal estimated likelihood function $\hat{L}_k(\theta_k \mid y) = L_k(\theta_k, \hat{\eta}_k \mid y)$ in step $k = 2, \ldots, K$. Here $\hat{\eta}_k$ is obtained from primary parameter estimates obtained in previous steps. The SMLE is detailed in Algorithm 1.

Algorithm 1 Sequential Maximum Likelihood Estimation

1: $\hat{\theta}_1 \leftarrow \arg \max_{\theta_1} L_1(\theta_1 \mid y)$
2: \textbf{for} $k \leftarrow 2$ to $K$ \textbf{do}
3: \hspace{1em} $\hat{\theta}_k \leftarrow \arg \max_{\theta_k} \hat{L}_k(\theta_k \mid y)$
4: \textbf{end for}
5: \textbf{return} $\hat{\theta} = \hat{\theta}_1 \cup \cdots \cup \hat{\theta}_K$

In general this is a sequential algorithm; under certain conditions, however, sequences of steps can be parallelized and performed in one stage. For example, the steps from $k + 1$ to $k + n$ can be parallelized if $\hat{\eta}_{k+j}$ for $j = 1, \ldots, n$ can be obtained after step $k$, since all the marginal estimated likelihood functions can be obtained after step $k$. Formally, the steps from $k + 1$ to $k + n$ can be performed independently if $\eta_{k+j} \subseteq \theta_1 \cup \cdots \cup \theta_k$ for $j = 1, \ldots, n$. Note that the SMLE method can be performed in parallel when estimating the parameters of the ARMA models of the diagonal VARMA model, see Subsection 3.3. Hence, the parameters of the diagonal VARMA model can be obtained in two stages.

3.2 Consistency

We assume the data set $y$ is a realization from a MP model with true parameter set $\theta^*$ and denote $Y$ as the corresponding random vector. Define $\hat{\theta}_1(Y)$ as the estimator of $\theta_1$, define $\hat{\theta}_k(Y, \eta_k)$ as the estimator of $\theta_k$ given $\eta_k$, define $\hat{\theta}_k'(Y, \eta_k)$—if it exists—as its
Jacobian matrix (derivative with respect to \( \eta_k \)) and define \( \hat{\eta}_k(Y) \) as the estimator of \( \eta_k \). Furthermore, let \( n_k \) quantify the information contained in \( Y \) relevant to the estimation of \( \theta_k \). For the diagonal VARMA model \( n_k = T \) for \( k \) from 1 to \( S \) and \( n_K = S \). The SMLE consistency theorem provides the conditions under which \( \hat{\theta}_k(Y, \hat{\eta}_k(Y)) \) is consistent for \( k = 2, \ldots, K \).

**Theorem 3.1 (SMLE Consistency).** Suppose that
\[
\hat{\theta}_1(Y) \xrightarrow{P} \theta_1^* \text{ as } n_1 \to \infty, \tag{2}
\]
\[
\hat{\theta}_k(Y, \eta_k^*) \xrightarrow{P} \theta_k^* \text{ as } n_k \to \infty, \tag{3}
\]
for \( k = 2, \ldots, K \) where \( \eta_k^* \) and \( \theta_k^* \) are the true parameter sets for all \( k \). Furthermore, assume that there exists a \( n_{k0} < \infty \) such that for all \( n_k > n_{k0} \), \( \hat{\theta}_k(Y, \eta_k) \) exists and is uniformly bounded in an open neighborhood of \( \eta_k^* \) almost surely for \( k = 2, \ldots, K \). Then
\[
\hat{\theta}_k(Y, \hat{\eta}_k(Y)) \xrightarrow{P} \theta_k^* \text{ as } n_1, \ldots, n_k \to \infty,
\]
for \( k = 2, \ldots, K \).

The assumptions of the SMLE consistency theorem are required for the Spall consistency theorem (Spall, 1989, Theorem 1) used in the inductive hypothesis of the proof provided in appendix B. Heuristically, for step \( k = 2, \ldots, K \) the theorem states that if \( \hat{\eta}_k \) is a consistent estimator, \( \hat{\theta}_k(Y, \eta_k^*) \) is a consistent estimator and \( \hat{\theta}_k'(Y, \eta_k) \) exists and is well-behaved near \( \eta_k^* \), then \( \hat{\theta}_k(Y, \hat{\eta}_k) \) is a consistent estimator.

### 3.3 Diagonal VARMA Inference

As \( \eta_k = \emptyset \) for \( k \) from 1 to \( S \), the parameters \( \theta_k \) for \( k \) from 1 to \( S \) can be estimated in parallel with ARMA likelihood functions in one stage. The temporal parameter estimates obtained in this stage are consistent in \( T \) (Hamilton, 1994, Section 5.8) and satisfy assumptions (2) and (3) of the SMLE consistency theorem (3.1) where \( n_k = T \) for \( k \) from 1 to \( S \). The step
$K$ marginal (joint) likelihood function has an innovation form (Schweppe, 1965)

$$\mathcal{L}_K(\theta_K, \eta_K \mid y) = \prod_{t=1}^{T} g(\theta_K \mid u_t(\eta_K \mid y)),$$  \hspace{1cm} (4)

where $g(\cdot)$ is the innovation likelihood function and $u_t(\eta_K \mid y)$ are the $S$ residuals of the data set at time point $t$ that depend on $\eta_K$. In stage two the spatial parameter estimates are obtained from (4) using the residuals obtained in stage one. Hence, this is a two-stage approach. This approach has some advantages. First, ARMA model selections can be performed in the first stage, i.e. the AR and MA orders of the ARMA models can vary across space. Second, innovation model selection and specification can be performed in the second stage. Note that if the innovation model is a MP model and the parameter set can be estimated in two stages, then the parameter set of the diagonal VARMA model can be estimated in three stages, see the application in Section 5 for an example.

4 Simulation Study

The first simulation study compares the small sample biases and standard errors of the maximum likelihood and stepwise maximum likelihood estimators for the diagonal VARMA model (1) introduced in Subsection 2.2 with isotropic innovations. The second simulation study is used to corroborate the SMLE consistency Theorem 3.1 introduced in Section 3.2 for the same model.

4.1 Simulation Model

The diagonal VARMA model (1) is used with zero mean, AR order two and MA order zero, i.e. centered diagonal VAR(2), with isotropic innovations. The number of parameters is restricted so that the SMLE method can be directly compared with the MLE method where the number of evaluations grows exponentially with the number of parameters. For all $T$ and $S$ considered in this section define the standard deviation and AR parameters as $\sigma_s = 1.2$, $\phi_{1,s} = 0.50$ and $\phi_{2,s} = 0.25$ for all $s$. The isotropic innovations are modeled with the Matérn correlation function, which for distance $h$ has the following form:
\[ \frac{\pi^{1/2}}{2^{\kappa-1} \Gamma(\kappa+1/2)\alpha^{2\kappa}} \left( \kappa K_\kappa(\alpha \| h \|) \right), \]

where the inverse scale parameter \( \alpha > 0 \) controls the range of correlation, the smoothness parameter \( \kappa > 0 \) controls the mean-square differentiability of the process and \( K_\kappa(\cdot) \) is a modified Bessel function (Stein, 1999, p. 31). In this section we define the inverse scale and smoothness parameters as \( \alpha = 0.3 \) and \( \kappa = 1.5 \) respectively. This model corresponds to \( N = T \cdot S \) data points, where spatial locations are distributed regularly on a line, and \( P = 3S + 2 \) parameters, i.e. three temporal parameters for each time series and two spatial parameters.

### 4.2 Set-up

Since, for MLE, the number of evaluations grows exponentially with the number of parameters and \( P = 3S + 2 \), the number of spatial locations \( S \) is restricted so that the SMLE method can be directly compared to the MLE method. Therefore, for the purpose of comparison, we let \( T = 50 \) and \( S = 20 \) (so that \( N = 50 \cdot 20 = 1,000 \) and \( P = 3 \cdot 20 + 2 = 62 \)). For both methods the algorithm is initialized at the true parameter values to eliminate the effects of initial value selection and standard errors are obtained by performing both the MLE and SMLE methods for 30 independent simulations. All numerical optimizations (for MLE and SMLE) are performed in R with the Nelder-Mead simplex algorithm (Nelder and Mead, 1965). For SMLE the first \( S \) steps are performed in parallel via hyper-threading with a 2.6 GHz Intel Core i7 (8 virtual cores) CPU and 8 GB of RAM.

### 4.3 Results

Table 1 displays the mean estimates (averaged over \( s \)) and standard errors, in parenthesis, of \( \sigma_s, \phi_{1,s}, \phi_{2,s}, \alpha \) and \( \kappa \) obtained from SMLE, Full MLE and Fixed MLE (\( \sigma_s \) fixed for all \( s \)) using 30 independent simulations. The estimates from Full MLE suggest that \( \sigma_s \) for all \( s \) are not identifiable in the diagonal VARMA model but are identifiable in the ARMA models. Consequently, Fixed MLE was included to provide a comparison with SMLE. The mean estimates and standard errors obtained from Fixed MLE aim to approximate those from Full MLE. The approximated estimates are expected to be more statistically efficient as fixing the standard deviation parameters has increased the data to parameter ratio. There is a relatively small difference in the sample biases of \( \hat{\phi}_{1,s} \) and \( \hat{\phi}_{2,s} \) between SMLE...
and Fixed MLE, however, the 15% and 14% relative efficiencies (ratio of standard errors), respectively, demonstrate that the estimates from SMLE are less efficient, as expected. The sample biases of $\hat{\alpha}$ and $\hat{\kappa}$ are relatively small with relative efficiencies 50% and 57% respectively. Considering that these are conservative estimates and the number of data points is small ($N = 1,000$) there appear to be relatively small sample biases. The fact that SMLE was capable of estimating $\sigma_s$ for all $s$ demonstrates one advantage of estimating with marginal likelihood functions.

Table 1: Mean estimates (averaged over $s$) and standard errors, in parenthesis, of $\sigma_s$, $\phi_{1,s}$, $\phi_{2,s}$, $\alpha$ and $\kappa$ obtained from SMLE, Full MLE and Fixed MLE using 30 independent simulations.

|                  | $\hat{\sigma}$ | $\hat{\phi}_1$ | $\hat{\phi}_2$ | $\hat{\alpha}$ | $\hat{\kappa}$ |
|------------------|----------------|----------------|----------------|----------------|----------------|
| SMLE             | 1.16 (0.12)    | 0.51 (0.13)    | 0.21 (0.14)    | 0.33 (0.04)    | 1.54 (0.07)    |
| Full MLE         | 95.13 (1746.27)| 0.30 (0.59)    | -0.03 (0.52)   | 0.15 (0.11)    | 1.27 (0.40)    |
| Fixed MLE        | NA             | 0.50 (0.02)    | 0.24 (0.02)    | 0.30 (0.02)    | 1.51 (0.04)    |
| True Values      | 1.20           | 0.50           | 0.25           | 0.30           | 1.50           |

Table 2 displays the mean time (seconds) and mean number of iterations required by both estimation methods and the two steps of the SMLE method. On average MLE took approximately 4,800 times longer (24 minutes 4 seconds) and required approximately 700 times more iterations than SMLE. The relative errors of the times and iterations are 6% and 10%, respectively, for SMLE and 34% and 24%, respectively, for MLE.

Estimators are expected to have zero bias and variance as $T$ and $S$ increase to infinity according to the consistency theorem. However, for clarity we provide plots of the bias, variance and MSE of estimators as $T$ increases with $S$ fixed then as $S$ increases with $T$ fixed. Figure 1 displays the bias (squared), variance and MSE profile plots of $\hat{\alpha}$ and $\hat{\kappa}$ from SMLE as the number of time points $T$ increases to 100 with $S = 20$ fixed and then as the number of spatial locations $S$ increases to 45 with $T = 100$ fixed. These plots provide numerical evidence of the SMLE consistency theorem in Subsection 3.2 as they demonstrate
Table 2: Mean time (seconds) and mean number of iterations required by both estimating methods and the two steps of the SMLE method. Means and standard deviations, in parenthesis, are calculated from the 30 simulations.

|                | SMLE   |               | MLE    |               |
|----------------|--------|---------------|--------|---------------|
|                | Step 1 | Step 2        | Total  | Total         |
| Time (sec)     | 0.09 (0.00) | 0.20 (0.02)  | 0.30 (0.02) | 1444.17 (486.67) |
| # Iterations   | 125.73 (18.12) | 50.60 (5.72)  | 176.33 (18.27) | 126178.30 (30820.19) |

how the MSE decreases for \( \hat{\alpha} \) and \( \hat{\kappa} \) conditional on the temporal parameter estimates as \( T \) increases for a fixed number of spatial locations \( (S = 20) \) and as \( S \) increases for a fixed number of time points \( (T = 100) \). For both \( \hat{\alpha} \) and \( \hat{\kappa} \) the variance dominates the bias for all values of \( T \). Furthermore, the plots demonstrate that the reductions in MSE are mostly attributable to reductions in variance with small bias even for a few time points and spatial locations.
Figure 1: Bias (squared), variance and mean square error (MSE) of $\hat{\alpha}$ and $\hat{\kappa}$ from SMLE as the number of time points $T$ increases to 100 with $S = 20$ fixed and then as the number of spatial locations $S$ increases to 45 with $T = 100$ fixed.

5 Application

5.1 Data

To illustrate the flexibility of a MP model and the efficiency of SMLE for large data we consider the large ensemble community project data set (Kay et al., 2015). This data set is an ensemble of Earth System Model simulations, consisting of hundreds of spatio-temporal variables. From this data set a single member of the near surface (2 m above surface) temperature variable is selected. This variable was simulated from 2006 to 2100...
according to the Representative Concentration Pathways 8.5 scenario (Van Vuuren et al., 2011), \( T = 95 \) years, over a discrete global grid consisting of \( N = 288 \) longitudes and \( M = 192 \) latitudes. The temperature variable was annually averaged to approximate, through the central limit theorem, the Gaussian assumption of the diagonal VARMA model (diagnostics of the other assumption is provided in Castruccio and Genton (2018)), for a total of \( T \times N \times M = 5,253,120 \) data points.

5.2 A Model for Global Data

A diagonal VAR(2) model is applied, and innovations \( U_t \) are assumed stationary in longitude but nonstationary in latitude (axial symmetry, Jones, 1963). The vector of mean parameters as in (1) is \( \mu = \beta_0 + t\beta_1 \) where \( \beta_j = (\beta_{1,j}, \ldots, \beta_{S,j}) \) for \( j = 1, 2 \). Given the discrete geometry of the data and the axial symmetry assumption, the correlation matrix \( R(\nu) \) of \( U_t \) in (1) is block-circulant (Davis, 1979), so the \( M \times M \) blocks are circulant with the form

\[
R_{m_1, m_2} = CF_{m_1, m_2} C^H, \quad m_1, m_2 = 1, \ldots, M,
\]

where \( C = C/\sqrt{N} \) is a unitary matrix, \( C \) is the unnormalized \( N \times N \) discrete Fourier transform (DFT) matrix and \( F_{m_1, m_2} = \text{diag}(f_{m_1, m_2}) \) is a diagonal matrix (Davis, 1979 Section 3.2). The elements of \( f_{m_1, m_2} \) are the values of the cross-spectral mass function, denoted \( f_{m_1, m_2}(\cdot) \), between latitudes \( L_{m_1} \) and \( L_{m_2} \). Hence, block-circulant matrices are modeled in the spectral domain. The cross-spectral mass functions have the form

\[
f_{m_1, m_2}(c) = f_{m_1}^{1/2}(c) \cdot f_{m_2}^{1/2}(c) \cdot \rho_{m_1, m_2}(c),
\]

where \( f_m(\cdot) \) is the spectral mass function at latitude \( L_m \) and \( \rho_{m_1, m_2}(\cdot) \) is the coherence function between latitudes \( L_{m_1} \) and \( L_{m_2} \).

Following Castruccio and Stein (2013); Poppick and Stein (2014) the spectral mass function for the innovations at latitude \( L_m \) is the modified Matérn defined as

\[
f_m(c) \propto \frac{1}{(\alpha_m^2 + 4 \sin^2 \frac{\pi c}{N})^{\kappa_m+1/2}} \quad \text{s.t.} \quad \frac{1}{N} \sum_{c=0}^{N-1} f_m(c) = 1,
\]
where $\alpha_m$ is the inverse range parameter and $\kappa_m$ controls the increased decay rate in spectral mass for larger wavenumbers. The normalization condition is required as the innovations are unscaled. The coherence function is defined as

$$
\rho_{m_1,m_2}(c) := \left[ \frac{\xi}{(1 + 4 \sin^2 \frac{\pi c}{N})^{\tau}} \right] |L_{m_1} - L_{m_2}|,
$$

where $\xi$ controls the rate of decay in coherence, over all wavenumbers, as the distance between latitudes increases and $\tau$ controls the increased decay rate in coherence for larger wavenumbers. Therefore, the correlation matrix $R(\nu)$ has the parameter set

$$
\nu = \left( \bigcup_{m=1}^{M} \alpha_m \right) \cup \left( \bigcup_{m=1}^{M} \kappa_m \right) \cup \xi \cup \tau.
$$

### 5.3 Estimation

The model presented for the innovations $U_t$ in the previous section is a MP model, and $\nu$ can be estimated in two stages. Consider the finite sequence of $K = M + 1$ data subsets $(u_m)$ where $u_m$ are the innovations at a fixed latitude $L_m$ for $m = 1, \ldots, M$ across longitudes (a circle), and $u_K = u$. The marginal model of $u_m$ is a Whittle model (Whittle, 1954) that depends on a parameter subset with a partition $\theta_m, \eta_m$ where $\theta_m = \{\alpha_m, \kappa_m\}$ and $\eta_m = \emptyset$, the proof is provided in Appendix C. The marginal (joint) model for $u_K$ depends on a parameter subset with a partition $\theta_K, \eta_K$ where $\theta_K = \{\xi, \tau\}$ and $\eta_K = \theta_1 \cup \cdots \cup \theta_M$. Clearly, this finite sequence of data subsets satisfies the conditions of Definition 1. Consequently, the parameter set of the diagonal VARMA model can be estimated in three stages rather than two in the case of an unstructured correlation matrix $R(\nu)$. In the first stage the temporal diagonal VAR parameters are estimated, in the second stage the longitudinal spectral mass function parameters are estimated and in the third stage the latitudinal coherence function parameters are estimated.

Figures 2 displays the mean, standard deviation, trend, and order one AR coefficient parameters of the marginal AR models over the discrete global grid. The mean parameters capture the variation in temperature that results from latitude and altitude (e.g. Himalayas, Andes). The standard deviation parameters capture the difference in temperature variabili-
ity over land surface and sea. The trend parameters capture the lower rates of increase of
temperature over the sea. The order one autoregressive parameters capture the long range
dependence over the equatorial Pacific ocean due to irregular periodic variation of the El Niño Southern Oscillation.

Figure 2: The mean (upper left), standard deviation (lower left), trend (upper right) and
order one AR coefficient (lower right) parameters of the marginal AR models over the
discrete global grid.

Figure 3 (top row) displays periodograms (blue) and fitted spectral mass functions (red)
at four selected latitudes. Figure 3 (bottom row) displays cross-periodograms (blue) and
fitted cross-spectral mass functions (red) at the corresponding latitudes. These figures
demonstrate that the spectral mass functions and the coherence function have provided a
reasonable model for the correlations within and between, respectively, the data at each
latitude given the axially symmetric assumptions.
Estimation was performed on the laptop computer described in Subsection 4.2. Table 3 displays the time required by the SMLE algorithm to estimate the temporal, longitudinal and latitudinal parameters. This table also includes the number of parameters and data points corresponding to each marginal model for the three stages. In total, this model has been fit to over five million data points in forty minutes. This model can be fit to larger data sets if a cluster of computers is available. However, the size and complexity of this data set demonstrates that fitting a diagonal VARMA model to a large complex spatio-temporal data set is possible on a laptop computer.

6 Conclusion and Future Work

This work lays the foundation of a new approach for modeling large and complex spatio-temporal data, by predicating a partition of the parameter space in order to achieve inference in a multi-step fashion, the key principle being that the dependence of some data subsets can be described exclusively by some parameters in the model.
Table 3: The time required by the SMLE algorithm to estimate the temporal, longitudinal, and latitudinal parameters, and the number of parameters and data points corresponding to each submodel for the three stages.

| Stage          | Time   | # Parameters / Marginal Model | # Data points / Marginal Model |
|----------------|--------|-------------------------------|-------------------------------|
| Temporal       | 9 mins | 4                             | 95                            |
| Longitudinal   | 7 secs | 2                             | 27,360                        |
| Latitudinal    | 31 mins| 3                             | 55,253,120                    |

MP models are very broad and flexible, and can be efficiently applied to large complex data sets via SMLE. The SMLE method can be parallelized, and we provide conditions under which the estimators are consistent. We demonstrate that a diagonal VARMA is MP, and that the parameters of this model can be estimated at least three orders of magnitude faster using SMLE compared to MLE, with very little loss in statistical efficiency. A MP model was applied to a large complex global climate data set consisting of over five million data points with a laptop computer in forty minutes. The model demonstrated the capacity to capture important complex features such as spatially varying means and nonstationary global dependence.

Proposing a MP model for a particular application requires the existence of a sequence of data subsets that satisfies the definition, and while in our application this was suggested by the geometry of the problem, an automatic approach comprising of a clustering method to identify this sequence would be desirable. It would also be desirable to develop a selection of MP models, for different applications, that satisfies the assumptions of the SMLE consistency theorem.

The parameter set of a MP model can be estimated efficiently via SMLE. This is a frequentist method that depends on the assumption that parameter subsets can be separated from the joint model with marginal models and estimated with marginal likelihood functions. Since Bayesian methods have been applied to composite marginal likelihood functions ([Pauli et al.] 2011) there could be the opportunity to embed MP models in tools.
of modern Bayesian inference.

The first simulation study suggested that the standard deviation parameters are not identifiable in the diagonal VARMA model but are identifiable in the ARMA models. In general, it is important to know if the identifiability of a set of parameters can change with respect to a marginal model. If the identifiability of a set of parameters can change with respect to a marginal model, it is still of interest to understand if this is the case for the diagonal VARMA model or whether these parameters are just very difficult to estimate.

References

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008), ‘Gaussian predictive process models for large spatial data sets’, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(4), 825–848.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012), ‘Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach’, Journal of the American Statistical Association 107(497), 268–280.

Castruccio, S. and Genton, M. G. (2016), ‘Compressing an Ensemble with Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature’, Technometrics 58(3), 319–328.

Castruccio, S. and Genton, M. G. (2018), ‘Principles for Inference on Big Spatio-Temporal Data from Climate Models’, Statistics and Probability Letters 136, 92–96.

Castruccio, S. and Guinness, J. (2017), ‘An Evolutionary Spectrum Approach to Incorporate Large-scale Geographical Descriptors on Global Processes’, Journal of the Royal Statistical Society: Series C (Applied Statistics) 66(2), 329–344.

Castruccio, S., Huser, R. and Genton, M. (2016), ‘High-order composite likelihood inference for max-stable distributions and processes’, Journal of Computational and Graphical Statistics 25(4), 1212–1229.
Castruccio, S., Ombao, H. and Genton, M. G. (2018), ‘A Scalable Multi-Resolution Spatio-
Temporal Model for Brain Activation and Connectivity in fMRI Data’, *Biometrics*. in
press.

Castruccio, S. and Stein, M. L. (2013), ‘Global space-time models for climate ensembles’,
*The Annals of Applied Statistics* 7(3), 1593–1611.

Cressie, N. and Johannesson, G. (2008), ‘Fixed rank kriging for very large spatial data sets’,
*Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(1), 209–
226.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016), ‘Hierarchical nearest-
neighbor Gaussian process models for large geostatistical datasets’, *Journal of the Amer-
ican Statistical Association* 111(514), 800–812.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., Schaap, M. et al. (2016), ‘Nonseparable
dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an
application to particulate matter analysis’, *The Annals of Applied Statistics* 10(3), 1286–
1316.

Davis, P. J. (1979), *Circulant Matrices*, American Mathematical Society, Providence, Rhode
Island.

Giltinan, D. and Davidian, M. (1995), *Nonlinear models for repeated measurement data*,
Chapman and Hall.

Golub, G. H. and Van Loan, C. F. (2012), *Matrix Computations*, Vol. 3, JHU Press.

Hamilton, J. D. (1994), *Time Series Analysis*, Vol. 2, Princeton University Press.

Johannesson, G., Cressie, N. and Huang, H.-C. (2007), ‘Dynamic multi-resolution spatial
models’, *Environmental and Ecological Statistics* 14(1), 5–25.

Jones, R. H. (1963), ‘Stochastic processes on a sphere’, *The Annals of mathematical statistics* 34(1), 213–218.
Katzfuss, M. (2017), ‘A multi-resolution approximation for massive spatial datasets’, *Journal of the American Statistical Association* **112**(517), 201–214.

Kaufman, C. G., Schervish, M. J. and Nychka, D. (2008), ‘Covariance tapering for likelihood-based estimation in large spatial data sets’, *Journal of the American Statistical Association* **103**(484), 1545–1555.

Kay, J., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S. C., Danabasoglu, G., Edwards, J. et al. (2015), ‘The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability’, *Bulletin of the American Meteorological Society* **96**(8), 1333–1349.

Lindgren, F., Rue, H. and Lindström, J. (2011), ‘An explicit link between Gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer Science & Business Media.

Nelder, J. A. and Mead, R. (1965), ‘A simplex method for function minimization’, *The Computer Journal* **7**(4), 308–313.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F. and Sain, S. (2015), ‘A multiresolution Gaussian process model for the analysis of large spatial datasets’, *Journal of Computational and Graphical Statistics* **24**(2), 579–599.

Pauli, F., Racugno, W. and Ventura, L. (2011), ‘Bayesian composite marginal likelihoods’, *Statistica Sinica* **21**(1), 149–164.

Pawitan, Y. (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press.

Poppick, A. and Stein, M. L. (2014), ‘Using covariates to model dependence in nonstationary, high-frequency meteorological processes’, *Environmetrics* **25**(5), 293–305.
Schabenberger, O. and Gotway, C. A. (2017), *Statistical Methods for Spatial Data Analysis*, CRC press.

Schweppe, F. C. (1965), ‘Evaluation of likelihood functions for Gaussian signals’, *IEEE Transactions on Information Theory* 11(1), 61–70.

Spall, J. C. (1989), ‘Effect of imprecisely known nuisance parameters on estimates of primary parameters’, *Communications in Statistics-Theory and Methods* 18(1), 219–237.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F. et al. (2011), ‘The representative concentration pathways: an overview’, *Climatic change* 109(1-2), 5.

Vavasis, S. A. (1991), *Nonlinear Optimization: Complexity Issues*, Oxford University Press, Inc.

Whittle, P. (1954), ‘On stationary processes in the plane’, *Biometrika* 41(3-4), 434–449.

Yudin, D. B. and Nemirovskii, A. S. (1977), ‘Evaluation of the informational complexity of mathematical programming problems’, *Matekon* 13(3), 25–45.

Yudin, D. B. and Nemirovskii, A. S. (1983), *Problem Complexity and Method Efficiency in Optimization*, Wiley.

Zhang, B., Sang, H. and Huang, J. Z. (2015), ‘Full-scale approximations of spatio-temporal covariance models for large datasets’, *Statistica Sinica* 25(1), 99–114.

**Appendix A  Marginal Diagonal VARMA Model**

*Proof.* If the diagonal VARMA model is stable then it can be represented as \(Y_t = \mu + \sum_{i=0}^{\infty} \Psi_i \Sigma U_{t-i}\) (Lütkepohl 2005, p. 421) where \(\Psi_i\) is a linear combination of the diagonal matrices of AR and MA parameters. Consequently, \(\Psi_i = \text{diag}(\psi_{i,1}, \ldots, \psi_{i,S})\)
are the diagonal matrices of infinite order MA parameters. The mean vector and lag-$h$ autocovariance matrix function of this model are $\mu$ and

$$\Gamma(h) := \sum_{i=0}^{\infty} \Psi_{i+h} \Sigma R(\nu) \Sigma \Psi_i$$

respectively. Therefore, the mean and lag-$h$ autocovariance function of the marginal model of $y_k$ is $\mu_k$ and $\gamma_k(h) := \sigma_k^2 \sum_{i=0}^{\infty} \psi_{i+h,k} \psi_{i,k}$. The marginal model of $y_k$ is therefore an ARMA model that depends on a parameter subset with a partition $\theta_k, \eta_k$ where

$$\theta_k = \mu_k \cup \sigma_k \cup \left( \bigcup_{i=1}^{\infty} \psi_{i,k} \right) = \mu_k \cup \sigma_k \cup \left( \bigcup_{i=1}^{p} \phi_{i,k} \right) \cup \left( \bigcup_{j=1}^{q} \pi_{j,k} \right)$$

and $\eta_k = \emptyset$ for $k = 1, \ldots, S$.

\section*{Appendix B \quad SMLE Consistency}

\textit{Proof.} Base case ($k = 2$): Since $\hat{\eta}_2(Y) \subseteq \hat{\theta}_2(Y)$ by definition, $\hat{\eta}_2(Y)$ is consistent by assumption and $\hat{\theta}_2(Y, \hat{\eta}_2(Y))$ is consistent by the Spall consistency theorem \cite{Spall1989}. 

Inductive hypothesis ($k > 2$): Suppose the theorem holds for $k = 2, \ldots, n < K$. Since $\hat{\eta}_{n+1}(Y) \subseteq \hat{\theta}_1(Y) \cup \cdots \cup \hat{\theta}_n(Y, \hat{\eta}_n(Y))$ by definition, $\hat{\eta}_{n+1}(Y)$ is consistent by the inductive hypothesis and $\hat{\theta}_{n+1}(Y, \hat{\eta}_{n+1}(Y))$ is consistent by the Spall consistency theorem \cite{Spall1989}.

\section*{Appendix C \quad Marginal Axially Symmetric Model}

\textit{Proof.} The block of the block-circulant correlation matrix that corresponds to $u_k$ is the circulant matrix $R_{k,k}$. Since $\rho_{k,k}(c) = 1$, the model has the following log probability density
function

\[ -\frac{1}{2} \ln |2\pi R_{k,k}| - \frac{1}{2} u_k^H R_{k,k}^{-1} u_k \]

\[ = -\frac{1}{2} \ln |2\pi U F_{k,k} U^H| - \frac{1}{2} (F u_k)^H F_{k,k}^{-1} F u_k \]

\[ = -\frac{1}{2} \ln |2\pi F_{k,k}| - \frac{1}{2} \tilde{u}_k F_{k,k}^{-1} \tilde{u}_k \]

\[ = -\frac{N}{2} \ln (2\pi) - \frac{1}{2} \sum_{c=0}^{N-1} \left( \ln f_k(c) - \frac{\bar{u}_{k,c}^2}{f_k(c)} \right) \]

where \( \tilde{u}_{k,c} \) is an element of \( \tilde{u}_k \). The marginal model of \( u_k \) is therefore a Whittle model (Whittle, 1954) that depends on a parameter subset with a partition \( \theta_k, \eta_k \) where \( \theta_k = \{ \alpha_k, \kappa_k \} \) and \( \eta_k = \emptyset \) for \( k = 1, \ldots, M \). \( \square \)