DISSERTATION

STATE-SPACE MODELS FOR STREAM NETWORKS

Submitted by

William J. Coar

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctorate of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2007

COLORADO STATE UNIVERSITY

March 29, 2007

WE HEREBY RECOMMEND THAT THE DISSERTATION STATE-SPACE MODELS FOR STREAM NETWORKS PREPARED UNDER OUR SUPERVISION BY WILLIAM J. COAR BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTORATE OF PHILOSOPHY.

Committee on Graduate Work

F. Jay Breidt (Adviser and Department Head)

Richard A. Davis (Committee Member)

Jennifer A. Hoeting (Committee Member)

Jose Salas (Committee Member)

ABSTRACT OF DISSERTATION

STATE-SPACE MODELS FOR STREAM NETWORKS

The natural branching that occurs in a stream network, in which two upstream reaches merge to create a new downstream reach, generates a tree structure. Furthermore, because of the natural flow of water in a stream network, characteristics of a downstream reach may depend on characteristics of upstream reaches. Since the flow of water from reach to reach provides a natural time-like ordering throughout the stream network, we propose a state-space model to describe the spatial dependence in this tree-like structure with ordering based on flow. This state-space model includes a state vector that evolves from reach to reach as a function of upstream reaches, and a measurement vector that depends on the state and allows for general spatial-temporal dependence among measurements on a reach.

Current methods of estimation and prediction on a stream network are based on Universal Kriging, where the covariance function is defined in terms of distance between measurement locations. However, because of the branching structure, the class of valid covariance functions becomes more restrictive than the general class available for spatially correlated data.

Application of a state-space model over other tree structures has been studied, but in a very different context. Areas such as multiscale resolution and Gaussian directed trees are similar topologically, but model assumptions for these networks are not always applicable to stream networks.

Developing a state-space formulation permits the use of the well known Kalman recursions. Variations of the Kalman Filter and Smoother are derived for the tree-structured state-space model, which allows recursive estimation of unobserved states

and prediction of missing observations on the network, as well as computation of the Gaussian likelihood, even when the data are incomplete. To reduce the computational burden that may be associated with optimization of this exact likelihood, a version of the expectation-maximization (EM) algorithm is presented that uses the Kalman smoother to fill in missing values in the E-step, and maximizes the Gaussian likelihood for the completed dataset in the M-step.

An example of a state-space model with parameters that depend on a surrogate for flow is presented. Simulation results for the exact likelihood, an EM algorithm, and a simplified EM algorithm are obtained. Maximum likelihood estimates and Monte Carlo standard errors for this two parameter estimation problem are presented.

Several forms of dependence for discrete processes on a stream network are considered, such as network analogues of the autoregressive-moving average model and stochastic trend models. Network parallels for first and second differences in time-series are defined, which allow for definition of a spline smoother on a stream network through a special case of a local linear trend model.

The methods developed here are applied to data available from Maryland's Department of Environmental Protection. A Moving Average is fit to a measure of instream cover in fish habitat data in a study that determines that autocorrelation can be removed by using appropriate spatial covariates. A smoothing spline is obtained to describe water chemistry data on this same network. Maximum Likelihood estimators are found for all unknown parameters.

The proposed models describe a discrete process, and can be used as a building block for continuous processes on a network. Adaptation of this state-space model and Kalman prediction equations to allow for more complicated forms of spatial and perhaps temporal dependence is a potential area of future research. Other possible directions for future research are non-Gaussian and non-linear error structures, model selection, and properties of estimators.

William J. Coar
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Spring 2007

ACKNOWLEDGEMENTS

The following people were instrumental toward completion of my dissertation...

To Jay Breidt, I am grateful for your willingness to share your experiences and knowledge, and providing continual guidance with patience through the numerous projects in which I have been involved. Your encouragement to always move just one step further has pushed me beyond my own expectations and has provided me with confidence in myself and my work. You have been a teacher and a role model, providing inspiration to strive and excel in the field of statistics.

To my committee, for your patience and willingness to provide direction when needed.

To my parents, for your continual support throughout this life changing experience. Your love and encouragement, although seldom acknowledged on my part, was vital. I am thankful.

To Cesar Kuroki and Lilliam Kingsbury, for being mentors and providing me with the experiences and encouragement to pursue this degree.

To the rest of my friends, both new and old. Over these years, I have experienced quite a change in my life while chasing this degree, this goal. I may have easily succumbed to the hard times and frustrations without having you to lean on. Each one of you has somehow been there, perhaps unknowingly, and this is something I will never forget.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Because of the natural flow of water in a stream network, characteristics of a downstream reach may depend on characteristics of upstream reaches, where a reach is defined as a section of stream between two confluences. Since the flow of water from reach to reach provides a natural time-like ordering throughout the stream network, we propose a state-space model to describe the spatial dependence in this tree-like structure with ordering based on flow. We use the inherent ordering based on flow as a tool in modeling dependence in a manner general enough to encompass a large class of stochastic processes which possess different forms of dependence.

Standard methods that define spatial correlation as a function of distance between two points have been considered for estimation and prediction on a stream network using geostatistical models (Peterson et al., 2006; Ver Hoef et al., 2006; Cressie et al., 2006). However, the tree structure of a stream network reduces the class of valid covariance functions (Peterson et al., 2006; Ver Hoef et al., 2006). There have also been state-space models developed for data on tree structures similar to stream networks that require both a forwards and backwards representation for prediction (Chou et al., 1994; Huang and Cressie, 2001; Tzeng et al., 2005). However, the natural evolution described by the corresponding forwards model produces inherent attributes that are not always applicable to a stream network. Specifically, given downstream information, upstream reaches would be considered independent if we were to consider this backwards model. The state-space formulation in this

dissertation fits within the spatial statistical structure of lattice data (discrete spatial support) rather than geostatistical data (continuous spatial support). It allows a stochastic process to evolve with flow, overcomes the need for both a forwards and backwards representation, and appropriately allows for conditional dependence of upstream reaches given downstream information.

1.1 Previous Work

Modeling dependence within a stream network has been done primarily through a function of distance between locations, as is typically done with geostatistical data, where observations "closer" in space tend to be more similar. Dependence is usually seen through a semi-variogram (Cressie, 1993, p.58). Monestiez et al. (2005) and Dent and Grimm (1999) directly fitted semi-variograms, but the results were dependent on bin size selected and did not guarantee variance constraints. Other geostatistical methods using for predicting along stream networks have been developed by Ver Hoef et al. (2006) and Cressie et al. (2006). Ver Hoef et al. (2006) and Cressie et al. (2006) develop methods to model dependence as a function of stream distance and Euclidean distance, and use kriging as a tool for prediction at unsampled sites and block kriging to estimate reach totals. Peterson et al. (2006) provide a review of studies exploring patterns of spatial autocorrelation in stream chemistry as functions of distance.

Dent and Grimm (1999) investigated spatial trends on a steam using one dimensional transects. Monestiez et al. (2005) used conditional probabilities defined by a directed tree to define spatial dependence on a stream network in terms of a curvilinear distance along the river. In these cases, weighted least squares was used to directly model the semi-variogram.

There are numerous autocovariance models commonly used to describe dependence as a function of distance between two points (see Cressie, 1993) that can be

considered for a stream network. More recently, stream distance has been considered as an alternative to straight line distance. Stream distance is the distance over which water must flow from one point to another. Tools such as Geographic Information Systems, commonly referred to as GIS, have made fairly accurate estimation of these distances possible. Ver Hoef et al. (2006) show that with proper weighting, many of the usual spatial models such as an exponential form of dependence can be used for stream networks, whereas others such as the spherical covariance model are invalid for the tree structure of a stream network (Ver Hoef et al., 2006; Peterson et al., 2006; Peterson and Urquhart, 2006).

Ver Hoef et al. (2006) develop a class of valid models that incorporate flow and stream distance by using spatial moving averages. These methods integrate a moving average function against a white noise process. The models incorporate flow by running the moving average function upstream, downstream, or both. The effect of different distance measures was investigated in Peterson et al. (2006). Once the distances are known and a covariance model selected, the final covariance matrix is easily constructed. This then allows kriging on a stream network once estimation of unknown parameters are obtained.

A weighted asymmetric hydrologic distance measure was also considered (Ver Hoef et al., 2006; Peterson et al., 2006; Peterson and Urquhart, 2006) in which locations are considered independent if water did not flow from one eventually to the other. A spatial weight matrix was obtained, where individual weights were defined by accumulating upstream catchment area if sites were flow connected, whereas the weight was set to zero otherwise.

Attempts have been made to incorporate both stream distance as well as straight-line distance. These methods were discussed at the Fourth Annual Conference in Statistics for Aquatic Resources - Monitoring, Modeling and Management (Oregon State University, Corvalis, OR, 2005). Cressie et al. (2006) developed kriging methods based on river network covariance functions developed by

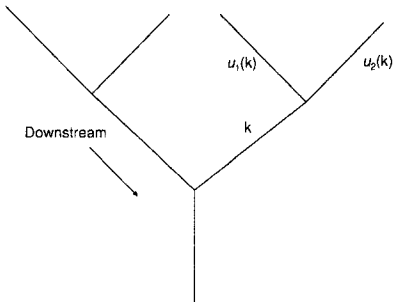


Figure 1.1: A simple example of a stream network consisting of seven reaches. The parents of reach k are denoted by $u_1(k)$ and $u_2(k)$.

Ver Hoef et al. (2006) for predicting change in dissolved oxygen content on a river network with sparse observation stations. Cressie et al. (2006) modeled dependence as a function of Euclidean distance as well as stream distance, and used ordinary kriging and a constrained kriging for prediction.

As an alternative approach to predicting a response observed in a stream network, the downstream flow and merging of streams depicted in Figures 1.1, 2.2, and 4.14 suggests the use of a state-space model to describe within-network dependence. Given a state-space form, commonly used tools such as the Kalman Filter (Kalman, 1960) could be used for estimation and prediction. State-space representations for other tree structures have already been developed (Chou et al., 1994; Basseville et al., 1992) in multiscale resolution problems and by Huang (1997), Huang and Cressie (2001) and Huang et al. (2002) in graphical models. These ideas were further enhanced for spatial prediction from large datasets involving change-of-resolution (Tzeng et al., 2005; Johannesson and Cressie, 2004; Johannesson et al., 2007).

The stochastic processes studied in these cases evolve over a tree from a single node to many terminal nodes, referred to as a forwards model. The state-space representations developed in these other tree structures closely resemble those used in

time series. Child nodes are created independently from parent nodes on these trees, so a state-space model only needs to evolve from one parent to each child individually, thus eliminating any obvious sign of a tree structure other than notation. Because of the parent-to-child evolution on these tree structures, a recursion that first moves up and then down the tree is needed to allow for measurement on the tree at one point to contribute to the estimate at another point on the same level of the tree. Thus, a backwards representation is required to move back up through the tree. Appropriate backwards models have developed (Chou et al., 1994; Huang and Cressie, 2001) based on on the work of Verghese and Kailath (1979), where the backwards model formulation was based on forwards model assumptions.

The backwards, or *reverse*, representation of a forward state-space formulation over a dyadic tree was used by Chou et al. (1994) in multiscale recursive estimation, such as reconstruction of noise degraded signals. Observations in multiscale representations have a natural time-like variable, *scale*. Processing signals at multiple scales involves pyramid-like structures, where each level in the pyramid corresponds to a particular scale and each node at a given scale is connected to both a parent node at the next coarser scale and several descendants at the next finer scale. The usual scale-to-scale resolution refinement by a factor of two leads directly to a binary tree structure.

A form of the Kalman Recursions was developed for such a stochastic process over a tree in Chou et al. (1994). Formulation of a fine-to-coarse prediction was developed from an appropriate backwards representation, whereas the coarse-to-fine smoothing operation was derived from the forward model. Under the assumed structure of the error terms in the forward model, the immediate descendants of a parent node are conditionally independent, given information from the parent. This is similar in form to the conditional independence assumed by Monestiez et al. (2005). Since each descendant is created independently, each can be used to obtain

a predicted value for the parent. Although the filtering algorithm involves the usual update step, the simultaneous predictions from both (or several) descendants of a parent are "fused" together to obtain a single prediction based on finer scale information. Conditional expectation and the joint distribution of the state of the parent with all the observed descendent information show the appropriate weighting to be a function of prediction error variances of each individual prediction from the child nodes. Details can be found in Chou et al. (1994, Appendix A).

The down-tree smoother proposed by Chou et al. (1994) smooths only one descendant at a time. The smoothed estimate of a node involves modifying the filtered estimate based on finer scale information with that from more coarse scales through the smoothed estimate at a nodes parents. The derivations are similar to fixed-interval smoothing in an ordinary time series. Details can be found in Chou et al. (1994, Appendix B).

The tree-structured models studied by Chou et al. (1994), a subclass of acyclic graphical models, can lead to blocky artifacts in predicted values in the same lineage. These blocks occur because of predictions at neighboring nodes come from different parents. Because of this, Huang (1997) proposed more general multiscale graphical models in terms of scale recursive dynamics on acyclic directed graphs, a generalization of one dimensional Markov chains where the only direct influence on a node is from its parents. A Bayesian approach to the derivation of a Kalman Filter for acyclic directed graph was presented, and further adapted to (Gaussian) undirected graphical models where the conditional distribution of a node conditioned on all other nodes depends only on its neighbors.

The Gaussian density and Bayes' theorem are used to derive the conditional distributions needed in both filtering and smoothing (Huang, 1997; Huang and Cressie, 2001). For the acyclic directed graph, "up-tree" filtering step at a particular node involves conditioning on the information from its children. This derivation relies

heavily on the assumption that conditioned on a parent node, the child nodes of that parent are independent. Moreover, this assumption also leads to simultaneous predictions for each parent node from each child. Using Bayes' theorem in this Gaussian case, the derived conditional expectation appropriately weights each individual prediction, as did the fusion of estimates presented by Chou et al. (1994). For the "down-tree" smoothing, the smoothed estimate is derived for each child individually, using the smoothed information from the parent node and the filtered information from the "up-tree" recursions.

This work has led to a variety of applications involving spatial prediction (Huang and Cressie, 2001; Huang et al., 2002; Tzeng et al., 2005) and change-of-resolution (Johannesson and Cressie, 2004; Johannesson et al., 2007). However, model assumptions pertaining to the parent-to-child evolution in these processes prevent further consideration.

1.2 Relation to Steam Networks

The geostatistical models have proven to be useful tools in prediction on stream networks. However, it has been shown that selection of an appropriate distance measure to correspond with ecological processes must be considered (Peterson et al., 2006). It has also been shown that the class of valid autocovariance functions on a stream network depends on the distance measure selected (Ver Hoef et al., 2006; Peterson et al., 2006; Peterson and Urquhart, 2006). After exploring the spatial dependencies associated with several water chemistry responses, Peterson et al. (2006) indicate that there is no clear answer of which to use, and suggest straight line distance because of simplicity since the models considered had similar predictive capabilities.

The state-space model we propose is not designed for geostatistical spatial data, but for lattice data. We adapt the class of Autoregressive Moving Average Models

as well as other stochastic trend models to obtain a large class of models for different forms of dependencies. In these models a process evolves from many initial nodes, or states, to a single terminal node. The state-space formulation will model dependence with respect to geometry. An advantage of the state-space form comes from adapted Kalman Recursions, which eliminates the need for construction of spatial weight matrices and inversion of large covariance matrices associated with larger networks.

The methods developed for tree structures such as multiscale resolution and directed trees may have application in the stream network setting, where the "down-stream" process is thought to be analogous to the backwards direction in a "coarse-to-fine" stochastic process. Although the ideas provide insight to methods for stream networks, model assumptions for multiscale resolution and directed trees prevent further consideration for stream networks. The backwards recursions on a binary tree start with the terminal nodes. For each set of two terminal nodes, updated predictions are used individually to make a prediction of the state at the parent node. These estimates are then merged with the necessary weighting to avoid double counting, a consequence of the conditional independence of the descendants. In order to utilize these recursions, it is necessary to assume that given the down-stream reach, the two upstream reaches are independent, an assumption that is unreasonable in many applications for stream networks.

The concepts of stationarity defined by Chou et al. (1994) are different than than those we consider in the stream network, where up-tree and down-tree transitions are used to define covariance. The work of Ver Hoef et al. (2006) considers a form of dependence similar in idea to that of Chou et al. (1994). We do not consider this type of dependence in this dissertation, where we model reaches as independent if water does not flow from one (eventually) to another. This idea is consistent with the asymmetric models considered by Ver Hoef et al. (2006) and Peterson et al. (2006).

1.3 Outline of Dissertation

In this dissertation, we adapt tools utilized in time series to the tree structure of a stream network. The outline is as follows. In Chapter 2 we define the state-space model and derive the Kalman recursions for the stream network. We further derive an alternative form of the recursions for cases when model assumptions allow for infinite variances associated with states of first order reaches.

We define the Gaussian Likelihood for data on a stream network in Chapter 3. For cases when model assumptions allow, we derive a concentrated likelihood, where a closed form expression is obtained for the initial mean vector associated with the states where water begins to flow. Resubstitution leads to a function of fewer parameters. When missing data are present, we provide an alternative state-space form for which the exact likelihood can be constructed. Furthermore, we define an Expectation-Maximization algorithm as an optimization tool for parameter estimation when the exact likelihood is complex.

In Chapter 4, we introduce the class of Autoregressive Moving Average (ARMA) Models for the tree structure of a stream network. As in the case of a time series ARMA model, we use the roots of the autoregressive polynomials to obtain parameter constraints ensuring second-order stationarity and expressions for the autocovariance function. We also provide a construction for the sample autocovariance and autocorrelation functions. We show a state-space form for these ARMA models, and provide a simulation example in which several different low order models with varying forms of dependence are considered.

We define stochastic trend models in Chapter 5. We provide network analogues of the Random Walk plus Noise (RW+N) and Local Linear Trend (LLT) models. We also use a special case of the LLT to define a discrete smoothing spline for the stream network.

We conclude with a brief summarization of our findings, and provide direction for future research.

Chapter 2

STATE-SPACE REPRESENTATION FOR A STREAM NETWORK

We consider a state-space representation and variation of a Kalman Filter and Kalman Smoother to assist in explaining the dependence within a stream network. We begin by ignoring distance between points therefore measuring dependence with respect to geometry rather than linear distance. There is no inherent time component in this model, as it is replaced by flow. To describe flow, we adopt the ordering developed by Strahler (1957), which is a simplified version of the commonly used indexing for river topologies originally introduced by Horton (1945) in the studies of river networks. Although this indexing itself possesses a multitude of mathematical properties, its use here is primarily to assist in describing flow through the network.

2.1 Notation

Due to the inconsistencies in hydrologic literature, we will use the following definitions for *reach* and *segment*. A reach is defined by a section of stream between two confluences, whereas a stream segment will be defined by a series of flow-connected reaches with a common order. Attempting to completely specify every reach in a network is notationally cumbersome. The identification of a particular reach of interest will be labeled by k. For convenience, generic subscripts u_1 and u_2 will denote parents of a reference reach k. If specific reference to the parent reaches are needed or not obvious, the parents will be denoted by $u_1(k)$ and $u_2(k)$. Any reference $j < k$ implies that j is *upstream* of k, or that k is a future generation of j. Likewise, $j > k$ implies that j is *downstream* of k.

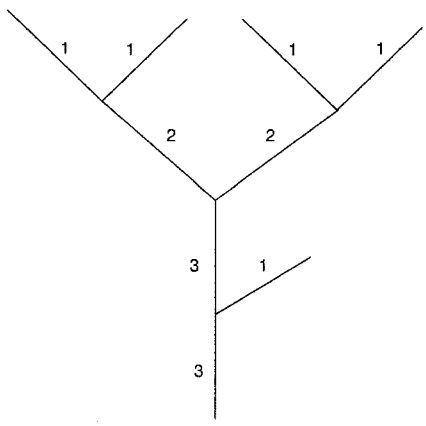


Figure 2.1: Example of a stream network identifing Strahler order of each reach. The order of the downstream reach is a function of the orders of the two immediate upstream reaches. The order increases by one if the two upstream orders are the same; otherwise, it is the maximum of the two upstream orders.

The indexing used to define the order, or rank, of a reach will be that developed by Strahler (1957) to describe progression through the network. By definition, the stream beds for initial drainages where water begins to flow are considered first order reaches. A recursive relationship then defines the order of any other reach k created by two parent sources, u_1 and u_2. The order of reach k, $r(k)$, is defined by the orders i, j of the two parent reaches: if $i \neq j$ then $r(k) = \max(i, j)$; otherwise $r(k) = i + 1$. See also Figure 2.1. A *higher* order reach is any reach downstream of a first order reach, where $r(k) > 1$. A stream *segment* of order k is a sequence of k^{th} order reaches connected by water flow starting when two $k-1$ order reaches merge and ending when a k^{th} order reach merges with another reach of order k or higher. Defined in this manner, stream segments are disjoint in that no two segments can share a common reach.

Matrices are generally identified by capital letters with subscripts, with the exception being cases where more than two indices are needed for matrix decomposition. Vectors will be displayed in bold with specific reference as well. For example, $\mathbf{X}(k)$ is a vector with components associated with reach k, and Ω_{u_1} is a matrix associated with one of the parents of reach k.

An additional summary table of notation adopted for this dissertation can be found in Appendix I.

2.2 The Autocovariance Function

The notion of *flow connected* is adopted when describing within-network dependence. If water from one reach (eventually) flows into another, then those two reaches are said to be flow connected. Reaches that are not flow connected are considered independent. This assumption has been made for other models of dependence within a stream network (Ver Hoef et al., 2006; Peterson et al., 2006; Peterson and Urquhart, 2006). Observations close in proximity will often be correlated, and the independence assumption would not hold in this case. However, if measurements are corrected for appropriate spatial covariates, this dependence can be removed, and the assumption is reasonable.

Modeling dependence throughout a stream network is done through an autocovariance function. For a network with a finite number of reaches, the autocovariance function is used to define the elements in the covariance matrix. Define $\mathcal{K}$ to be an index set used to identify any reach in the network, with possibly an infinite number of reaches. Although we refer to a particular reach by location, k is actually a one-dimensional indexing label. Then without loss in generality, we can define $\mathcal{K}$ to be $\mathbb{Z}$.

Definition 2.2.1 If $\{X(k), k \in \mathcal{K}\}$ is a process such that $\text{Var}\,(X(k)) < \infty$ for every $k \in \mathcal{K}$, then the autocovariance function $\gamma_X(\cdot,\cdot)$ of $\{\mathbf{X}(k)\}$ is defined by

$$\begin{aligned}\gamma_X(k,k') &= \text{Cov}(X(k), X(k')) \\ &= E\left[(X(k) - EX(k))(X(k') - EX(k'))\right], \quad k, k' \in \mathcal{K}.\end{aligned}$$

where $\gamma_X(k,k') \equiv 0$ when k and k' are not flow connected.

We now define (weak) stationarity, implying that the first and second moments of the process are independent of location k.

Definition 2.2.2 A process $\{X(k), k \in \mathcal{K}\}$ over index set $\mathcal{K}$ is said to be stationary if

(i) $E|X(k)|^2 < \infty$

(ii) $EX(k) = m$ for all $k \in \mathcal{K}$

(iii) $\gamma_X(k, k') \equiv \begin{cases} \gamma_X(h) & \text{if } k \text{ and } k' \text{ are flow connected} \\ 0 & \text{otherwise} \end{cases}$

where

$$h \equiv |k - k'| \equiv \text{the number of confluences between } k \text{ and } k'.$$

We will often refer to h as "lag" between two reaches.

Remark 2.2.1 The above definitions are analogous to those in time series defined by Brockwell and Davis (1991, Section 1.3) and elsewhere, where the indexing is no longer over time, but an arbitrary set of locations in space. Furthermore, as the covariance of a stationary time series is only a function of lag and independent of time t, stationarity on a stream network implies that dependence is strictly a function of flow connectivity and number of intermediate confluences.

As defined, this modeling on a stream network describes autocovariance between reaches due to the flow of water from one reach to the next. We model that data as if it were collected instantaneously, ignoring the time of collection.

2.3 State-Space Model

The observation equation expresses an observation vector, $\mathbf{Y}(k)$, as a linear function of the underlying state plus noise. Define the observation (or measurement)

equation by

$$\mathbf{Y}(k) = G_k\mathbf{X}(k) + \mathbf{W}(k) \tag{2.1}$$

where $\{G_k\}$ is a sequence of matrices with dimension that can depend on reach k and $\{\mathbf{W}(k)\}$ is an independent sequence of random vectors with mean zero and positive definite covariance matrix $\{R_k\}$.

In defining the state (or process) equation for the tree-like structure of the stream network, we recognize that any reach k is created by two upstream parents, with the exception of first order reaches, those in which water begins to flow. Initialization begins by defining an appropriate model for these first order reaches. The state $\mathbf{X}(k)$ at higher order reach k can then be expressed as a linear combination of both parents plus noise. This is written as

$$\mathbf{X}(k) = F_{k,u_1}\mathbf{X}(u_1) + F_{k,u_2}\mathbf{X}(u_2) + \mathbf{V}(k) \tag{2.2}$$

where F_{k,u_i} defines the contribution from parent $u_i(k)$, and $\{\mathbf{V}(k)\}$ is a sequence of independent zero-mean random vectors each with positive definite covariance matrix $\{Q_k\}$. Although it is not a restriction of the model, we initially consider Gaussian error structures in both the observation and state equations. We further assume that all first order states $\mathbf{X}(k)$ are multivariate Normal with mean $\mathbf{a}_k$ and covariance Ω_k, either of which may be unknown. We note that this model assumes the state is constant over the entire reach, and transitions only occur at a confluence with another reach.

The state-space model adopted for the stream networks is very general in that the system matrices can be dependent on location within the network. In particular, this allows for the dimensions to change, necessary for cases with unequal numbers of observations on different reaches. For example, a monitoring station may produce a regular time series at one location, while other sites are visited irregularly.

The state-space models under this formulation can describe both stationary and non-stationary processes on a stream network, as well as a combination thereof.

Because two merging reaches are not flow connected, they are assumed to be unconditionally independent, as was discussed in our definition of the autocovariance function. For some processes, this assumption may seem unrealistic unconditionally, since one would expect dependence due to their close proximity (in space). However, such dependence can often be removed with the appropriate covariates such as landscape characteristics. These covariates can either enter the model in the system matrices, or the state-space formulation can be used to model the residuals resulting from a different model that has accounted for the necessary covariates.

2.3.1 Lag, Differencing and the Backshift Operator

Because distance is ignored and dependence is associated with geometry, upstream parents as well as any other reaches further upstream can be identified in terms of *lag*. Define the $lag = |k - k'|$ in the stream network to be the number of reaches upstream that k is from k' with lag=0 corresponding to reach k. Lag can also be thought of as the number of generations between k and k'. In a binary tree, there are 2^i reaches at each lag i that eventually merge to create reach k. Let the subscript ij identify reach j at lag i from a reference reach k.

An operation commonly used in time series in order to obtain stationarity from a nonstationary process involves differencing between increments. On a tree structure, an increment can be defined by moving from one generation to the next. Applying this concept to a stream network, each increment is defined by two parent reaches merging at a confluence to create the downstream reach. Thus, differencing involves both parents.

The difference operator ∇ on a stream network is defined by

$$\nabla X(k) = X(k) - \frac{(X(u_1) + X(u_2))}{2} = (1 - B)X(k)$$

where the backshift operation B is defined by

$$BX(k) = \frac{X(u_1) + X(u_2)}{2}.$$

Powers of the backshift operator make use of lag notation previously described, with $B^i X(k) = \sum_{j=1}^{2^i} \frac{X_{ij}(k)}{2^i}$, which is an average of all 2^i ancestral reaches from the i^{th} upstream generation. Powers of ∇ and functions of B operate as any other polynomial function of a real variable. For example,

$$\begin{aligned}
\nabla^3 X(k) &= (1-B)^3 X(k) \\
&= (1 - 3B + 3B^2 - B^3) X(k) \\
&= X(k) - \frac{3}{2}\big(X_{11}(k) + X_{12}(k)\big) + \frac{3}{4}\big(X_{21}(k) + \cdots + X_{24}(k)\big) \\
&\quad -\frac{1}{8}\big(X_{31}(k) + \cdots + X_{38}(k)\big)
\end{aligned}$$

which can also be derived by $\nabla(\nabla^2)X(k)$.

2.4 Kalman Recursions

The Kalman Filter has proven to be a useful tool in providing inference about state vectors of time series, and has been generalized to other data structures with some inherent ordering, such as multi-scale data (Chou et al., 1994) and acyclic graphical models (Huang, 1997). The Kalman filter is an estimation procedure such that the optimal predictor of the current state vector is a linear combination of the optimal predictor at the previous state and the current observation. This allows predictions to be continually updated as observations become available. When future observations are obtained, the Kalman smoother provides a way of further updating the predictions using this added information. Since each of these procedures involve computation in steps, inversion of a large covariance matrix is avoided when working with large amounts of data. Furthermore, these procedures provide tools for estimation with or without completely observed data.

Before proceeding, we list the initial assumptions used in the analysis of the state-space models defined by (2.1) and (2.2). We continue to assume independence and Normality throughout, however, this can be replaced by orthogonality of random vectors and the results remain valid.

(a) F_k and G_k are specified matrices whose dimension may depend on $k \in \mathcal{K}$.

(b) $\mathbf{W}(k)$ and $\mathbf{V}(k)$ are independent Normal random vectors.

(c) For any first order reach k, $\mathbf{X}(k)$ is independent to $\mathbf{W}(k')$ and $\mathbf{V}(k')$, $k, k' \in \mathcal{K}$.

(d) $\mathbf{X}(k)$ is independent of $\mathbf{X}(k')$ if k and k' are not flow connected.

(e) $E\mathbf{V}(k) = \mathbf{0}$ and $E\mathbf{W}(k) = \mathbf{0}$ for all k.

(f) $E\left(\mathbf{V}(k)\mathbf{V}(k)^T\right) = Q_k$ and $E\left(\mathbf{W}(k)\mathbf{W}(k)^T\right) = R_k$.

Beginning with the additional assumption that $\mathbf{a}_k$ and Ω_k are known for all first order reaches, a variation of the Kalman recursions is derived. The Kalman prediction equations are obtained for a stream network. These equations define the Kalman filter, a two step process outlined in the following Theorem to obtain a prediction for the state on reach k based on observations on or upstream of k.

Theorem 2.4.1 *(Kalman Filter)*

For the state-space model defined by (2.1) and (2.2) with assumptions listed above, the prediction step in the Kalman Filter is defined by

$$\mathbf{X}^p(k) = F_{k,u_1}\mathbf{X}^f(u_1) + F_{k,u_2}\mathbf{X}^f(u_2) \tag{2.3}$$

with prediction error variance

$$\Omega_k^p = F_{k,u_1}\Omega_{u_1}^f F_{k,u_1}^T + F_{k,u_2}\Omega_{u_2}^f F_{k,u_2}^T + Q_k. \tag{2.4}$$

The prediction in (2.3) is optimal in the sense that $E\left(\mathbf{X}(k) - \hat{\mathbf{X}(k)}\right)^2$ *is minimized among all predictors* $\hat{\mathbf{X}(k)}$ *given the upstream information. These predictions and variances are then updated by*

$$\mathbf{X}^f(k) = \mathbf{X}^p(k) + \Omega_k^p G_k^T \Delta_k^{-1} \left(\mathbf{Y}(k) - G_k \mathbf{X}^p(k)\right) \tag{2.5}$$

with prediction error variance

$$\Omega_k^f = \Omega_k^p - \Omega_k^p G_k^T \Delta_k^{-1} G_k \Omega_k^p. \tag{2.6}$$

where Δ_k^{-1} *is any generalized inverse of* $\Delta_k = G_k \Omega_k^p G_k^T + R_k$. *This new prediction defined by (2.5) minimizes* $E\left(\mathbf{X}(k) - \hat{\mathbf{X}(k)}\right)^2$ *among all predictors* $\hat{\mathbf{X}(k)}$ *given the available information on or upstream of* k.

Remark 2.4.1 In the Gaussian case, the minimum mean square predictors for $\mathbf{X}(k)$ given the available data are defined by (2.3) and (2.5). Since conditional expectation is linear in the Gaussian case, these predictions are also the *best linear predictors* given the available data at each step.

Estimation is performed recursively, progressing with the flow of water via Strahler order and reach-within-segment order. Stream segments progress downstream with Strahler order, and reaches within a particular segment are naturally ordered by flow. The recursions begin with second order segments, after predicted and filtered values are obtained for first order streams through underlying initial conditions. Predicted and filtered estimates are calculated by reach-to-reach recursions within each segment, a process which is repeated for all segments, in a sequence based on segment (Strahler) order.

To obtain an estimate based on both upstream and downstream information, filtered values are modified two at a time with downstream information to obtain smoothed predictions, an iterative process defined in the following Theorem.

Theorem 2.4.2 *(Kalman Smoother)*

Smoothed estimates for upstream parent reaches are defined by

$$\begin{bmatrix} \mathbf{X}^s(u_1) \\ \mathbf{X}^s(u_2) \end{bmatrix} = \begin{bmatrix} \mathbf{X}^f(u_1) \\ \mathbf{X}^f(u_2) \end{bmatrix} + \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} (\mathbf{X}^s(k) - \mathbf{X}^p(k)) \tag{2.7}$$

where $\Theta_{u_i,k} = \Omega^f_{u_i} F^T_{k,u_i} (\Omega^p_k)^{-1}$. *The smoothed prediction* $\mathbf{X}^s(u_i)$ *minimizes* $E\left(\mathbf{X}(u_i) - \hat{\mathbf{X}(u_i)}\right)^2$ *among all predictors* $\hat{\mathbf{X}(u_i)}$ *given the available information on the stream network. The corresponding prediction error variance of a smoothed estimate is formulated to be*

$$\Omega^s_{u_i} = \Omega^f_{u_i} + \Theta_{u_i,k}(\Omega^s_k - \Omega^p_k)\Theta^T_{u_i,k} \tag{2.8}$$

with a cross covariance of

$$\Omega^s_{u_i,u_j} = \Theta_{u_i,k}(\Omega^s_k - \Omega^p_k)\Theta^T_{u_j,k}. \tag{2.9}$$

Remark 2.4.2 Similar to the case of the filtered predictions, the minimum mean square predictor for $\mathbf{X}(u_i)$ given all the available data, both upstream and downstream, is defined by (2.7). Since conditional expectation is linear in the Gaussian case, these predictions are also the *best linear predictors* given the available data on the stream network.

Similar to fixed-interval smoothing in time series, predicted and filtered values, as well as corresponding variances, are obtained for every reach in the network. Smoothing starts with the reach furthest downstream, smoothing the two parent reaches simultaneously, proceeding upstream until all first order reaches have been smoothed. In this upstream process, we simultaneously smooth both parent reaches, which by definition are on different segments. Because of the natural branching structure and order defined by Strahler, it is possible to smooth "up" each segment of a particular order until all segments of that order have been smoothed before proceeding to the next (sequentially) lower Strahler order.

In summary, the initial prediction, which is based solely on the upstream observations, is modified once new information on the current reach is available. This filtered estimate is further modified once downstream data are obtained. Together, Theorems (2.4.1) and (2.4.2) provide an estimate based on all available information on the stream network. Proofs of each Theorem follow.

2.4.1 Derivation of the Downstream Filter

The derivation of the Kalman filter involves two procedures, predicting and updating. Define $\mathcal{U}_{(k)}$ to be the set of reaches upstream of k, and $\mathcal{U}_{(k]} = \{k\} \cup \mathcal{U}_{(k)}$. The initial step involves predicting the $\mathbf{X}(k)$ based on information upstream of reach k, which is the observed data for $k' \in \mathcal{U}_{(k)}$. The second step involves updating the prediction based on new information from the observation at reach k, thus predicting $\mathbf{X}(k)$ given data for $k' \in \mathcal{U}_{(k]}$.

The Kalman filter in the context of the stream network given the assumed state-space representation with Gaussian noise terms consists of predictions and updates defined by

$$\begin{aligned} \mathbf{X}^p(k) &= E[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}] \\ \mathbf{X}^f(k) &= E[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k]}]. \end{aligned}$$

Consider the joint distribution of the three vectors $\mathbf{X}(k)$, $\mathbf{Y}(u_1)$ and $\mathbf{Y}(u_2)$ conditioned on $\{\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}\}$. Under the Normality assumptions, we see the conditional distribution

$$\left[\begin{array}{c} \mathbf{Y}(u_1) \\ \mathbf{Y}(u_2) \\ \mathbf{X}(k) \end{array} \right] \Bigg| \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}$$

is

$$\mathbf{N}\left(\left[\begin{array}{c} G_{u_1}\mathbf{X}^p(u_1) \\ G_{u_2}\mathbf{X}^p(u_2) \\ \mathbf{X}^p_{u_1,u_2}(k) \end{array} \right], \left[\begin{array}{ccc} \Delta_{u_1} & \mathbf{0} & G_{u_1}\Omega^p_{u_1}F^T_{k,u_1} \\ \mathbf{0} & \Delta_{u_2} & G_{u_2}\Omega^p_{u_2}F^T_{k,u_2} \\ F_{k,u_1}\Omega^p_{u_1}G^T_{u_1} & F_{k,u_2}\Omega^p_{u_2}G^T_{u_2} & \Omega^p_{u_1,u_2} \end{array} \right] \right)$$

where

$$\mathbf{X}^p_{u_1,u_2} = F_{k,u_1}\mathbf{X}^p(u_1) + F_{k,u_2}\mathbf{X}^p(u_2),$$

$$\Delta_{u_i} = G_{u_i}\Omega^p_{u_i}G^T_{u_i} + R_{u_i},$$

and

$$\Omega^p_{u_1,u_2} = F_{k,u_1}\Omega^p_{u_1}F^T_{k,u_1} + F_{k,u_2}\Omega^p_{u_2}F^T_{k,u_2} + Q_k.$$

Using conditional expectation with the Multivariate Normal distribution (Hocking, 1996, p.42), we find that

$$\begin{aligned}
\mathbf{X}^p(\mathbf{k}) &= E[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}] \\
&= E[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(u_1]} \cup \mathcal{U}_{(u_2]}] \\
&= E[\mathbf{X}(k)|\mathbf{Y}(u_1), \mathbf{Y}(u_2), \mathbf{Y}(k') \; \forall \; k' \in \mathcal{U}_{(u_1)} \cup \mathcal{U}_{(u_2)}] \\
&= F_{k,u_1}\mathbf{X}^p(u_1) + F_{k,u_1}\Omega^p_{u_1}G^T_{u_1}\Delta^{-1}_{u_1}\left(\mathbf{Y}(u_1) - G_{u_1}\mathbf{X}^p(u_1)\right) \\
&\quad + F_{k,u_2}\mathbf{X}^p(u_2) + F_{k,u_2}\Omega^p_{u_2}G^T_{u_2}\Delta^{-1}_{u_2}\left(\mathbf{Y}(u_2) - G_{u_2}\mathbf{X}^p(u_2)\right) \qquad (2.10)
\end{aligned}$$

with prediction error variance

$$\begin{aligned}
\Omega^p_k &= V[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}] \\
&= \Omega^p_{u_1,u_2} - \begin{bmatrix} F_{k,u_1}\Omega^p_{u_1}G^T_{u_1} & F_{k,u_2}\Omega^p_{u_2}G^T_{u_2} \end{bmatrix} \begin{bmatrix} \Delta^{-1}_{u_1} & 0 \\ 0 & \Delta^{-1}_{u_2} \end{bmatrix} \begin{bmatrix} G_{u_1}\Omega^p_{u_1}F^T_{k,u_1} \\ G_{u_2}\Omega^p_{u_2}F^T_{k,u_2} \end{bmatrix} \\
&= F_{k,u_1}\Omega^p_{u_1}F^T_{k,u_1} + F_{k,u_2}\Omega^p_{u_2}F^T_{k,u_2} + Q_k \\
&\quad - F_{k,u_1}\Omega^p_{u_1}G^T_{u_1}\Delta^{-1}_{u_1}G_{u_1}\Omega^p_{u_1}F^T_{k,u_1} - F_{k,u_2}\Omega^p_{u_2}G^T_{u_2}\Delta^{-1}_{u_2}G_{u_2}\Omega^p_{u_2}F^T_{k,u_2} \\
&= F_{k,u_1}\left(\Omega^p_{u_1} - \Omega^p_{u_1}G^T_{u_1}\Delta^{-1}_{u_1}G_{u_1}\Omega^p_{u_1}\right)F^T_{k,u_1} \\
&\quad + F_{k,u_2}\left(\Omega^p_{u_2} - \Omega^p_{u_2}G^T_{u_2}\Delta^{-1}_{u_2}G_{u_2}\Omega^p_{u_2}\right)F^T_{k,u_2} + Q_k \qquad (2.11)
\end{aligned}$$

where $\Delta^{-1}_{u_i}$ is a generalized inverse of Δ_{u_i}. Since conditional expectation minimizes mean square error in the Gaussian case, the predictions in (2.10) are optimal (Brockwell and Davis, 1991, p.64).

Once $\mathbf{Y}(k)$ is observed, the predicted values and variances are updated to obtain filtered values. If we consider the joint distribution of $\mathbf{Y}(k)$ and $\mathbf{X}(k)$ conditioned

on $\{\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}\}$,

$$\left[\begin{array}{c} \mathbf{Y}(k) \\ \mathbf{X}(k) \end{array} \right] \Big| \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)} \sim \mathbf{N}\left(\left[\begin{array}{c} G_k \mathbf{X}^p(k) \\ \mathbf{X}^p(k) \end{array} \right], \left[\begin{array}{cc} \Delta_k & G_k \Omega_k^p \\ \Omega_k^p G_k^T & \Omega_k^p \end{array} \right] \right)$$

and again use the conditional expectation, the predictions are updated by

$$\begin{aligned} \mathbf{X}^f(k) &= E[\mathbf{X}(k) | \mathbf{Y}(k), \mathbf{Y}(k') \; \forall \; k' \in \mathcal{U}_{(k)}] \\ &= E[\mathbf{X}(k) | \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]}] \\ &= \mathbf{X}^p(k) + \Omega_k^p G_k^T \Delta_k^{-1} \left(\mathbf{Y}(k) - G_k \mathbf{X}^p(k) \right) \end{aligned} \tag{2.12}$$

using the added information from $\mathbf{Y}(k)$. This new prediction minimizes the mean square error with this new information. The filtering variance associated with this update is

$$\Omega_k^f = \Omega_k^p - \Omega_k^p G_k^T \Delta_k^{-1} G_k \Omega_k^p. \tag{2.13}$$

In terms of this new formulation, we see that

$$\mathbf{X}^p(k) = F_{k,u_1} \mathbf{X}^f(u_1) + F_{k,u_2} \mathbf{X}^f(u_2) \tag{2.14}$$

$$\Omega_k^p = F_{k,u_1} \Omega_{u_1}^f F_{k,u_1}^T + F_{k,u_2} \Omega_{u_2}^f F_{k,u_2}^T + Q_k, \tag{2.15}$$

simplifying equations (2.10) and (2.11).

2.4.2 Derivation of the Upstream Smoother

The intent of the smoother is to predict $\mathbf{X}(k)$ based on all observed information. In the spirit of fixed-interval smoothing, the Kalman filter is first run over the network, saving the results from the forward recursions. Since each reach is created from the two upstream parents, the algorithm will smooth upstream two reaches at a time, starting with the reach furthest downstream.

The filtered estimate for reach k already uses information from reach k and above. The derivation of the smoother modifies this filtered estimate using the

information from downstream. Define $\mathcal{D}_{(k)}$ to be the set of all reaches downstream (hence flow connected) of k, with $\mathcal{D}_{[k)} = \{k\} \cup \mathcal{D}_{(k)}$. Then by definition, it is clear that $\{\mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)}\}$ consists of all the observed data used to obtain a smoothed estimate for $\mathbf{X}(k)$.

To begin, consider the conditional joint distribution of $\mathbf{X}(u_1)$, $\mathbf{X}(u_2)$ and $\mathbf{X}(k)$,

$$\left[\begin{array}{c} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \\ \mathbf{X}(k) \end{array} \right] \Bigg| \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}$$

which is

$$\mathbf{N}\left(\left[\begin{array}{c} \mathbf{X}^f(u_1) \\ \mathbf{X}^f(u_2) \\ \mathbf{X}^p(k) \end{array} \right], \left[\begin{array}{ccc} \Omega^f_{u_1} & \mathbf{0} & \Omega^f_{u_1} F^T_{k,u_1} \\ \mathbf{0} & \Omega^f_{u_2} & \Omega^f_{u_2} F^T_{k,u_2} \\ F_{k,u_1}\Omega^f_{u_1} & F_{k,u_2}\Omega^f_{u_2} & \Omega^p_k \end{array} \right] \right)$$

under the Normality assumptions used thus far. Further conditioning on $\mathbf{X}(k)$, we find

$$E\left[\begin{array}{c} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{array} \Bigg| \mathbf{X}(k), \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)} \right] = \left[\begin{array}{c} \mathbf{X}^f(u_1) \\ \mathbf{X}^f(u_2) \end{array} \right] + \left[\begin{array}{c} \Omega^f_{u_1} F^T_{k,u_1} \\ \Omega^f_{u_2} F^T_{k,u_2} \end{array} \right] (\Omega^p_k)^{-1} \left(\mathbf{X}(k) - \mathbf{X}^p(k)\right)$$

Since any $\mathbf{Y}(k')$ for k' downstream of k can be expressed as a function of $\mathbf{X}(k)$, independent error terms, and incoming processes independent of $\mathbf{X}(k)$, we see the conditional expectation

$$E\left[\begin{array}{c} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{array} \Bigg| \mathbf{X}(k), \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right]$$

is equivalent to

$$E\left[\begin{array}{c} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{array} \Bigg| \mathbf{X}(k), \mathcal{X}_k, \mathbf{Y}(k'), \mathbf{W}(k'), \mathbf{V}(k') \; \forall \; k' \in \mathcal{D}_{[k)} \right]$$

where $\mathcal{X}_k$ is the set of all downstream incoming processes that are not flow connected to k. From this we see that

$$E\left[\begin{array}{c} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{array} \Bigg| \mathbf{X}(k), \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right] = E\left[\begin{array}{c} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{array} \Bigg| \mathbf{X}(k), \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)} \right]$$

Using this relation, the derivation of the smoothed estimates for $\mathbf{X}(u_1)$ and $\mathbf{X}(u_2)$ is

$$\begin{aligned}
\begin{bmatrix} \mathbf{X}^s(u_1) \\ \mathbf{X}^s(u_2) \end{bmatrix} &= E\left[\begin{bmatrix} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{bmatrix} \middle| \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right] \\
&= E\left[E\left[\begin{bmatrix} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{bmatrix} \middle| \mathbf{X}(k), \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right] \middle| \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right] \\
&= E\left[E\left[\begin{bmatrix} \mathbf{X}(u_1) \\ \mathbf{X}(u_2) \end{bmatrix} \middle| \mathbf{X}(k), \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)} \right] \middle| \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right] \\
&= \begin{bmatrix} \mathbf{X}^f(u_1) \\ \mathbf{X}^f(u_2) \end{bmatrix} + \\
&\quad \begin{bmatrix} \Omega^f_{u_1} F^T_{k,u_1} \\ \Omega^f_{u_2} F^T_{k,u_2} \end{bmatrix} (\Omega^p_k)^{-1} E\left[\mathbf{X}(k) - \mathbf{X}^p(k) | \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)} \right] \\
&= \begin{bmatrix} \mathbf{X}^f(u_1) \\ \mathbf{X}^f(u_2) \end{bmatrix} + \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} (\mathbf{X}^s(k) - \mathbf{X}^p(k)) . \qquad (2.16)
\end{aligned}$$

where $\Theta_{u_i,k} = \Omega^f_{u_i} F^T_{k,u_i} (\Omega^p_k)^{-1}$. In cases where Ω^p_k is singular, any generalized inverse can be used. The smoothed predictions are said to be optimal in that they minimize the mean square error conditioned on all the observations (Brockwell and Davis, 1991, p.64).

The derivation of the smoothing variances $\Omega^s_{u_i}$ and covariance Ω_{u_1,u_2} uses orthogonality of a residual vector when conditioning on components of the Multivariate Normal distribution (see Hocking, 1996, p.44). The estimate in (2.12) is a conditional expectation, conditioned on $\{\mathbf{Y}(k'), k' \in \mathcal{U}_{(k]}\}$. Hence, the residual vector $\mathbf{X}(k) - \mathbf{X}^f(k)$ is independent of any $\mathbf{Y}(k')$ for $k' \in \mathcal{U}_{(k]}$. Similarly, (2.16) is also a conditional mean, conditioned on $\{\mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)}\}$, so the residual vector $\mathbf{X}(k) - \mathbf{X}^s(k)$ is independent of any $\mathbf{Y}(k')$ for $k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)}$.

Subtracting both sides of (2.16) from $\begin{bmatrix} \mathbf{X}(u_1)^T & \mathbf{X}(u_2)^T \end{bmatrix}^T$ and performing some algebraic manipulations, we see that

$$\begin{bmatrix} \mathbf{X}(u_1) - \mathbf{X}^s(u_1) \\ \mathbf{X}(u_2) - \mathbf{X}^s(u_2) \end{bmatrix} + \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} \mathbf{X}^s(k) = \begin{bmatrix} \mathbf{X}(u_1) - \mathbf{X}^f(u_1) \\ \mathbf{X}(u_2) - \mathbf{X}^f(u_2) \end{bmatrix} + \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} \mathbf{X}^p(k).$$

Working towards

$$\mathrm{E}[(\mathrm{LHS})(\mathrm{LHS})^T] = \mathrm{E}[(\mathrm{RHS})(\mathrm{RHS})^T],$$

the left-hand side (LHS) and right-hand side (RHS) are addressed separately, starting with the LHS first.

Since $\mathbf{X}^s(k)$ is a linear combination of $\mathbf{Y}(k'))$ for $k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)}$, then $(\mathbf{X}(k) - \mathbf{X}^s(k)) \perp \mathbf{X}^s(k)$, so the LHS results in

$$\begin{bmatrix} \Omega^s_{u_1} & \Omega^s_{u_1,u_2} \\ \Omega^s_{u_2,u_1} & \Omega^s_{u_2} \end{bmatrix} + \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} E\left[\mathbf{X}^s(k)\mathbf{X}^s(k)^T\right] \begin{bmatrix} \Theta^T_{u_1,k} & \Theta^T_{u_2,k} \end{bmatrix} \tag{2.17}$$

when the expectation is carried through. Similarly, orthogonality will allow simplification when the same operation is applied to the RHS. Note that $\mathbf{X}^p(k), \mathbf{X}^f(u_1)$, and $\mathbf{X}^f(u_2)$ are linear combinations of $\mathbf{Y}(k')$ for $k' \in \mathcal{U}_{(k)}$, so $(\mathbf{X}(u_i) - \mathbf{X}^f(u_i)) \perp \mathbf{X}^p(k)$ leaving the RHS to be

$$\begin{bmatrix} \Omega^f_{u_1} & 0 \\ 0 & \Omega^f_{u_2} \end{bmatrix} + \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} E\left[\mathbf{X}^p(k)\mathbf{X}^p(k)^T\right] \begin{bmatrix} \Theta^T_{u_1,k} & \Theta^T_{u_2,k} \end{bmatrix}. \tag{2.18}$$

Since (2.17) equals (2.18), we have that the smoothing matrix is

$$\begin{aligned} \begin{bmatrix} \Omega^s_{u_1} & \Omega^s_{u_1,u_2} \\ \Omega^s_{u_2,u_1} & \Omega^s_{u_2} \end{bmatrix} &= \begin{bmatrix} \Omega^f_{u_1} & 0 \\ 0 & \Omega^f_{u_2} \end{bmatrix} + \\ &\begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} E\left[-\mathbf{X}^s(k)\mathbf{X}^s(k)^T + \mathbf{X}^p(k)\mathbf{X}^p(k)^T\right] \begin{bmatrix} \Theta^T_{u_1,k} & \Theta^T_{u_2,k} \end{bmatrix}. \end{aligned} \tag{2.19}$$

Re-expressing the terms inside the expectation, we have

$$\begin{aligned} &\mathbf{X}(k)\mathbf{X}(k)^T - (\mathbf{X}^s(k) - \mathbf{X}(k) + \mathbf{X}(k))\mathbf{X}^s(k)^T \\ &-\mathbf{X}(k)\mathbf{X}(k)^T + (\mathbf{X}^p(k) - \mathbf{X}(k) + \mathbf{X}(k))\mathbf{X}^p(k)^T. \end{aligned}$$

Using orthogonality, this expectation results in

$$E\left[\mathbf{X}(k)(\mathbf{X}(k) - \mathbf{X}^s(k))^T - \mathbf{X}(k)(\mathbf{X}(k) - \mathbf{X}^p(k))^T\right] = \Omega^s_k - \Omega^p_k$$

which can be substituted back into (2.19) to obtain the desired result

$$\begin{bmatrix} \Omega^s_{u_1} & \Omega^s_{u_1,u_2} \\ \Omega^s_{u_2,u_1} & \Omega^s_{u_2} \end{bmatrix} = \begin{bmatrix} \Omega^f_{u_1} & 0 \\ 0 & \Omega^f_{u_2} \end{bmatrix}$$

$$+ \begin{bmatrix} \Theta_{u_1,k} \\ \Theta_{u_2,k} \end{bmatrix} (\Omega_k^s - \Omega_k^p) \begin{bmatrix} \Theta_{u_1,k}^T & \Theta_{u_2,k}^T \end{bmatrix}. \qquad (2.20)$$

From (2.20), it is seen that each smoothed vector $\mathbf{X}^s(u_i)$ has a smoothing variance of

$$\Omega_{u_i}^s = \Omega_{u_i}^f + \Theta_{u_i,k}(\Omega_k^s - \Omega_k^p)\Theta_{u_i,k}^T.$$

The cross covariance between the smoothed vectors $\mathbf{X}^s(u_1)$ and $\mathbf{X}^s(u_2)$ is

$$\Omega_{u_i,u_j}^s = \Theta_{u_i,k}(\Omega_k^s - \Omega_k^p)\Theta_{u_j,k}^T.$$

The type of cross dependence, either positive or negative, relies on F_{k,u_1} and F_{k,u_2}.

2.4.3 Alternative Backwards Recursive Relationship

The upstream, or backwards, recursive relationship defined in (2.16) can be re-expressed in order to formulate an adaptation of the Kalman recursions presented by Durbin and Koopman (2001, Chapter 5) for cases when initial conditions are unknown. Furthermore, notation introduced in this backwards relationship will also be seen in the development of a concentrated likelihood when initial states are unknown.

By definition, the smoothed estimate for the state $\mathbf{X}(k)$ is

$$\begin{aligned} \mathbf{X}^s(k) &= E\left[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)}\right] \\ &= E\left[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)} \cup \mathcal{D}_{[k)}\right] \\ &= E\left[\mathbf{X}(k)|\mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}, \mathbf{v}(k''), k'' \in \mathcal{D}_{[k)}\right] \end{aligned}$$

where the innovation $\mathbf{v}(k)$ is defined by $\mathbf{v}(k) = \mathbf{y}(k) - G_k\mathbf{X}^p(k)$. Using standard regression theory and properties of the Multivariate Normal distribution, we see that

$$\mathbf{X}^s(k) = \mathbf{X}^p(k) + \sum_{k' \in \mathcal{D}_{[k)}} \mathrm{Cov}\left(\mathbf{X}(k), \mathbf{v}(k')\right) \Delta_{k'}^{-1} \mathbf{v}(k')$$

from which it is clear that the smoothed estimate is the sum of $\mathbf{X}^p(k)$ and a linear combination of the innovations $\mathbf{v}(k'), k' \geq k$.

The weights of the downstream innovations can be determined via derivation of the covariances, but can also be seen by considering the first few backwards recursions. Starting with the furthest downstream reach in the network, the smoothed estimate is simply the filtered value,

$$\mathbf{X}^s(k) = \mathbf{X}^p(k) + \Omega_k^p G_k^T \Delta_k^{-1} \mathbf{v}(k). \tag{2.21}$$

Defining $\mathbf{r}(k) = G_k^T \Delta_k^{-1} \mathbf{v}(k)$ when k is the furthest downstream reach, substitution of (2.14) and (2.15) into (2.16) to smooth one of the first two upstream reaches, we see that

$$\begin{aligned}
\mathbf{X}^s(u_i) &= \mathbf{X}^f(u_i) + \Omega_{u_i}^f F_{k,u_i}^T (\Omega_k^p)^{-1} \left(\mathbf{X}^s(k) - \mathbf{X}^p(k)\right) \\
&= \mathbf{X}^p(u_i) + \Omega_{u_i}^p G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) \\
&\quad + \left(\Omega_{u_i}^p - \Omega_{u_i}^p G_{u_i}^T \Delta_{u_i}^{-1} G_{u_i} \Omega_{u_i}^p\right) F_{k,u_i}^T (\Omega_k^p)^{-1} \left(\Omega_k^p G_k^T \Delta_k^{-1} \mathbf{v}(k)\right) \\
&= \mathbf{X}^p(u_i) + \Omega_{u_i}^p G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) + \Omega_{u_i}^p L_{k,u_i}^T G_k^T \Delta_k^{-1} \mathbf{v}(k) \\
&= \mathbf{X}^p(u_i) + \Omega_{u_i}^p \left(G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) + L_{k,u_i}^T G_k^T \Delta_k^{-1} \mathbf{v}(k)\right) \\
&= \mathbf{X}^p(u_i) + \Omega_{u_i}^p \left(G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) + L_{k,u_i}^T \mathbf{r}(k)\right) \\
&= \mathbf{X}^p(u_i) + \Omega_{u_i}^p \mathbf{r}(u_i)
\end{aligned} \tag{2.22}$$

where $L_{k,u_i} = F_{k,u_i} - F_{k,u_i} \Omega_{u_i}^p G_{u_i}^T \Delta_{u_i}^{-1} G_{u_i}$ and $\mathbf{r}(u_i) = G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) + L_{k,u_i}^T \mathbf{r}(k)$. From this, we can see the recursive relationship in $\mathbf{r}(k)$, taking note of the definition of $\mathbf{r}(k)$ when k is the furthest downstream reach.

Regression theory can also be used to show a recursive relationship in the smoothed variance. Since

$$\begin{aligned}
\Omega_{u_i}^s &= \mathrm{Var}\left[\mathbf{X}(k) | \mathbf{Y}(k'), k' \in \mathcal{U}_{(k]} \cup \mathcal{D}_{(k)}\right] \\
&= \mathrm{Var}\left[\mathbf{X}(k) | \mathbf{Y}(k'), k' \in \mathcal{U}_{(k)}, \mathbf{v}(k''), k'' \in \mathcal{D}_{[k)}\right]
\end{aligned}$$

we can see that

$$\Omega^s_{u_i} = \Omega^p_{u_i} - \sum_{k' \in \mathcal{D}_{[k)}} \mathrm{Cov}\left(\mathbf{X}(k), \mathbf{v}(k')\right) \Delta^{-1}_{k'} \mathrm{Cov}\left(\mathbf{X}(k), \mathbf{v}(k')\right)^T$$

The backwards relationship can again be seen by considering the first few recursions. Starting with the furthest downstream reach, the smoothed variance is the filtered variance, defined by

$$\Omega^s_k = \Omega^p_k - \Omega^p_k G^T_k \Delta^{-1}_k G_k \Omega^p_k.$$

Using (2.20) and letting $N_k = G^T_k \Delta^{-1}_k G_k$ (recalling reach k is the furthest downstream), we see that

$$\begin{aligned}
\Omega^s_{u_i} &= \Omega^f_{u_i} + \Theta_{u_i,k}(\Omega^s_k - \Omega^p_k)\Theta^T_{u_i,k} \\
&= \Omega^f_{u_i} - \Omega^f_{u_i} F^T_{k,u_i} G^T_k \Delta^{-1}_k G_k F_{k,u_i} \Omega^f_{u_i} \\
&= \Omega^p_{u_i} - \Omega^p_{u_i} G^T_{u_i} \Delta^{-1}_{u_i} G_{u_i} \Omega^p_{u_i} \\
&\quad -\Omega^p_{u_i}(I - G^T_{u_i} \Delta^{-1}_{u_i} G_{u_i} \Omega^p_{u_i}) F^T_{k,u_i} N_k F_{k,u_i} (I - \Omega^p_{u_i} G^T_{u_i} \Delta^{-1}_{u_i} G_{u_i}) \Omega^p_{u_i} \\
&= \Omega^p_{u_i} - \Omega^p_{u_i} G^T_{u_i} \Delta^{-1}_{u_i} G_{u_i} \Omega^p_{u_i} - \Omega^p_{u_i} L^T_{k,u_i} N_k L_{k,u_i} \Omega^p_{u_i} \\
&= \Omega^p_{u_i} - \Omega^p_{u_i} N_{u_i} \Omega^p_{u_i}
\end{aligned}$$

where $N_{u_i} = G^T_{u_i} \Delta^{-1}_{u_i} G_{u_i} + L^T_{k,u_i} N_k L_{k,u_i}$.

2.4.4 Non-Gaussian Structure

Although the Kalman recursions for the stream network were derived using a Gaussian error structure and properties of the Multivariate Normal distribution, the Gaussian assumptions are not a restriction to either the model or the resulting prediction equations. Assuming finite second moments, the optimal predictor in mean square sense is defined by conditional expectation. In the Gaussian case, this conditional expectation is linear. However, in the Non-Gaussian case, the optimal predictor is generally non-linear. In this case, the Kalman equations defined by Theorems (2.4.1) and (2.4.2) result in the best *linear* predictor in mean square sense. The arguement for optimality among linear predictors is standard, and hence is omitted.

2.4.5 Matrix Notation

The following matrix representations will be used in deriving a concentrated likelihood under conditions to be addressed later, but the general idea is to decompose the innovation $(\mathbf{Y}(k) - G_k\mathbf{X}^p(k))$ into two components: one being a linear combination of initial state predictions of first order reaches upstream of k while the other is a linear combination of the observations upstream of k. Note that indices in matrix notation may identify the position of a matrix or vector, not necessarily a row or column. The development of the matrix representation follows that in Durbin and Koopman (2001, p.95) for time series, but matrices on a river network will have a block-like structure depending on a reach's specific upstream first order reaches.

The matrix form of the observation equation is relatively straightforward, expressed by

$$\mathbf{Y} = G\mathbf{X} + \mathbf{W}, \quad \mathbf{W} \sim N(\mathbf{0}, R) \tag{2.23}$$

where $\mathbf{Y}$ results from stacking $\mathbf{Y}(k)$. When these vectors are ordered based on flow through Strahler order and reach-within-segment, the block-like structure is more easily seen.

The block-like structures are inherent in the matrix form of the state equation. First notice that any higher order reach can be written as a linear combination of first order reaches, or initial states, upstream of that reach plus state noise. Let $\mathbf{X}_0$ be a stacked vector of initial states. If $\mathbf{V}(k) = \mathbf{0}$ for any first order k, then

$$\mathbf{X} \;=\; F\mathbf{X}_0 + S\mathbf{V}, \tag{2.24}$$

where S and F have a block-like lower triangular structure. The diagonal elements of F are identity matrices, I_{n_k} where n_k is the dimension of $\mathbf{X}(k)$. If j identifies

the reach corresponding to the j^{th} vector in $\mathbf{X}_0$, then the lower triangular blocks for reach k are defined by

$$F_{kj} = \begin{cases} \prod_{k' \in \mathcal{I}_{(j,k]}} F_{k',u(k')} & \text{if the } j^{th} \text{ vector in } \mathbf{X}_0 \text{ is upstream of } k, \\ 0 & \text{otherwise} \end{cases}$$

where $u(k')$ is the appropriate parent of k' that connects j with k. The matrix S has a slightly more complicated structure since $V(k)$ are not introduced with first order k, only with higher order reaches. Hence, the product is over a subset of intermediary reaches $\mathcal{I}_{(k',k]}$, specifically, the reaches $\mathcal{I}_{(k',k]}$ such that $r(k') > 1$. Call this set $\mathcal{H}_{(k',k]}$. Further note that $\mathbf{V}$ is a stacked vector of $\mathbf{V}(k)$, for all k. Then if j now identifies the reach corresponding to the j^{th} vector $\mathbf{V}$, we define

$$S_{kj} = \begin{cases} \prod_{k' \in \mathcal{H}_{(k(j),k]}} F_{k',u(k')} & \text{if } k(j) \text{ is upstream of } k, \\ 0 & \text{otherwise} \end{cases}$$

which is similar in form to the matrix F. From this we see that F has blocks of zeros in columns associated with first order reaches that are not ancestral to the k^{th} reach of interest, whereas S has blocks of zeros in columns associated with reaches that are not upstream of k. Variation in $\mathbf{V}$ only comes from $\mathbf{V}(k)$ for higher order k since these error terms are not introduced until a merge between two reaches occurs. Let Q be a matrix that defines the variation in $S\mathbf{V}$.

To obtain the decomposition in $\mathbf{v}(k)$, observe from (2.3) that

$$\begin{aligned} \mathbf{X}^p(k) &= F_{k,u_1}\mathbf{X}^f(u_1) + F_{k,u_2}\mathbf{X}^f(u_2) \\ &= L_{k,u_1}\mathbf{X}^p(u_1) + L_{k,u_2}\mathbf{X}^p(u_2) + K_{u_1}\mathbf{Y}(u_1) + K_{u_2}\mathbf{Y}(u_2) \end{aligned}$$

where $L_{k,u_i} = F_{k,u_i} - F_{k,u_i}\Omega^p_{u_i}G^T_{u_i}\Delta^{-1}_{u_i}G_{u_i}$ and $K_{u_i} = F_{k,u_i}\Omega^p_{u_i}G^T_{u_i}\Delta^{-1}_{u_i}$. Recursive substitution for upstream k can be used to identify the weights associated with each element in $\mathbf{X}_0$ as well as those for each $\mathbf{Y}(k)$. For first order reach j with $\mathbf{X}^p(j)$ in $\mathbf{X}_0$, the corresponding weight in predicting $\mathbf{X}(k), j < k$ is

$$\prod_{k' \in \mathcal{I}_{(j,k]}} L_{k',u(k')}.$$

and the weight associated with each upstream $Y(j)$ is

$$\left(\prod_{k' \in \mathcal{I}_{(j,u(k))}} L_{k',u(k')} \right) K_j.$$

The general form of $\mathbf{X}^p(k)$ can be seen as

$$\mathbf{X}^p(k) = \sum_{j \in \mathcal{F}_k} \left(\prod_{k' \in \mathcal{I}_{(j,k]}} L_{k',u(k')} \right) \mathbf{X}^p(j) + \sum_{j \in \mathcal{U}_{(k)}} \left(\prod_{k' \in \mathcal{I}_{(j,u(k))}} L_{k',u(k')} \right) K_j \mathbf{Y}(j)$$

from which it can be shown that $E[\mathbf{v}_k] = E[\mathbf{Y}(k) - G_k \mathbf{X}^p(k)] = \mathbf{0}$. In matrix notation, define K and L to be matrices consisting of the above weights. Note that L consists of weights associated with upstream first order initial states and K has the weights for all upstream observations. In matrix form, we see that

$$\begin{aligned} \mathbf{v} &= (I - GK)\mathbf{y} - GL\mathbf{X}_0 \\ &= C^*\mathbf{y} - GL\mathbf{X}_0. \end{aligned} \tag{2.25}$$

Since $E[\mathbf{V}] = \mathbf{0}$, we see that $C^* = GL$. Moreover, it is easy to show that C^* is a lower triangular matrix, and for a fixed $\mathbf{X}_0$, is used to obtain a Cholesky decomposition for the variance in $\mathbf{Y}$. The matrices K and L also have the (lower triangular) block-like structure since the weights for predicting $\mathbf{X}(k)$ are zero for any reach that is not upstream of k. Similar to the matrix F, each non-zero element in row k of L and K results from a product of matrices corresponding to reaches that connect upstream k' with k.

Example 2.4.1 The block-like triangular structures described in the above matrices are easily seen through a simple example. Consider a small stream network depicted in Figure 2.2. Here we define several of the matrices previously discussed.

For the observation equation 2.23, we have the vectors

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}(1) \\ \mathbf{Y}(2) \\ \mathbf{Y}(3) \\ \mathbf{Y}(4) \\ \mathbf{Y}(5) \\ \mathbf{Y}(6) \\ \mathbf{Y}(7) \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}(1) \\ \mathbf{X}(2) \\ \mathbf{X}(3) \\ \mathbf{X}(4) \\ \mathbf{X}(5) \\ \mathbf{X}(6) \\ \mathbf{X}(7) \end{bmatrix}$$

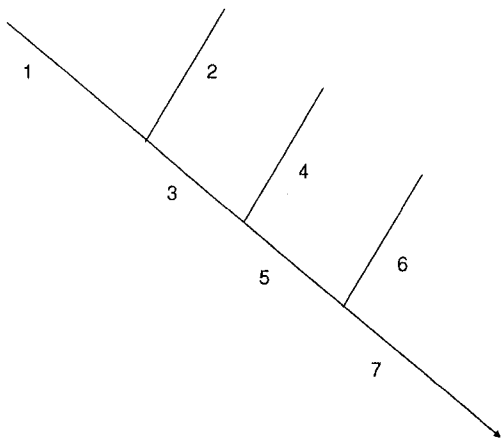


Figure 2.2: Example of a stream network consisting of seven reaches.

and diagonal matrices

$$G = \text{diag}\{G(k)\}_{k=1...7} \quad W = \text{diag}\{\mathbf{W}(k)\}_{k=1...7} \ .$$

We begin to see block-like structures in the components of the state equation (2.24). The vectors $\mathbf{X}_0$ and $\mathbf{V}$ are

$$\mathbf{X}_0 = \begin{bmatrix} \mathbf{X}(1) \\ \mathbf{X}(2) \\ \mathbf{X}(4) \\ \mathbf{X}(6) \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{V}(3) \\ \mathbf{0} \\ \mathbf{V}(5) \\ \mathbf{0} \\ \mathbf{X}(7) \end{bmatrix}$$

since $\mathbf{V}(k) = 0$ for first order k. The matrix F is

$$F = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ F_{3,1} & F_{3,2} & 0 & 0 \\ 0 & 0 & I & 0 \\ F_{5,3}F_{3,1} & F_{5,3}F_{3,2} & F_{5,4} & 0 \\ 0 & 0 & 0 & I \\ F_{7,5}F_{5,3}F_{3,1} & F_{7,5}F_{5,3}F_{3,2} & F_{7,5}F_{5,4} & F_{7,6} \end{bmatrix}$$

where S is

$$S = \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & F_{5,3} & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & F_{7,5}F_{5,3} & 0 & F_{7,5} & 0 & I \end{bmatrix}$$

For the decomposition in (2.25), we have

$$K = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ K_1 & K_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ L_{5,3}K_1 & L_{5,3}K_2 & K_3 & K_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ L_{7,5}L_{5,3}K_1 & L_{7,5}L_{5,3}K_2 & L7,5K_3 & L7,5K_4 & K_5 & K_6 & 0 \end{bmatrix}$$

where we recall that $K_{u_i} = F_{k,u_i}\Omega^p_{u_i}G^T_{u_i}\Delta^{-1}_{u_i}$. Lastly, with $L_{k,u_i} = F_{k,u_i} - F_{k,u_i}\Omega^p_{u_i}G^T_{u_i}\Delta^{-1}_{u_i}G_{u_i}$, we have

$$L = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ L_{3,1} & L_{3,2} & 0 & 0 \\ 0 & 0 & I & 0 \\ L_{5,3}L_{3,1} & L_{5,3}L_{3,2} & L_{5,4} & 0 \\ 0 & 0 & 0 & I \\ L_{7,5}L_{5,3}L_{3,1} & L_{7,5}L_{5,3}L_{3,2} & L_{7,5}L_{5,4} & L_{7,6} \end{bmatrix}.$$

The matrix C^* is easily obtained from the relation $C^* = I - GK = GL$.

2.5 Initialization

The recursions previously derived assume known initial conditions, specifically, that for any first order reach k, $\mathbf{X}(k) \sim \mathbf{N}(\mathbf{a}_k, \Omega_k)$ is normally distributed with known mean vector and covariance matrix. In many practical applications, these initial conditions will not be known. Durbin and Koopman (2001, Chapter 5) suggest a modification to the Kalman recursions to account for this, which we now adapt

to the recursions developed for a stream network. Suppose the m-dimensional state for any first order reach is of the form

$$\mathbf{X}(k) \;=\; \mathbf{a} + A\delta + V_0\eta_0, \quad \eta_0 \sim N(\mathbf{0}, Q_0)$$

where the $m \times 1$ vector $\mathbf{a}$ is known, δ is a $q \times 1$ vector of unknown quantities, the $m \times q$ matrix A and $m \times (m-q)$ matrix V_0 are matrices that consist of columns of the identity matrix $I_{m\times m}$such that $A^T V_0 = 0$. The matrices A and V_0 are selection matrices, as they are used to identify, or select, particular elements from the initial state vector. The matrix Q_0 is assumed positive definite and known. If $\mathbf{X}(k)$ is stationary, $A = 0$ and all elements in Q_0 can be derived from the model parameters. Essentially, this decomposes the initial state into a constant, a non-stationary component, and a stationary component.

Although the vector δ can be treated as fixed but unknown and estimated via maximum likelihood, we treat it as a Normal random vector with infinite variance defined by

$$\delta \sim \mathbf{N}(\mathbf{0}, \kappa I_q)$$

where we eventually let $\kappa \to \infty$. For first order reach k, we see that $E(\mathbf{X}(k)) = \mathbf{0}$ with variance Ω_k which has the form

$$\Omega_k = \kappa\Omega_{\infty,k} + \Omega_{*,k}$$

with $\Omega_{\infty,k} = AA^T$ and $\Omega_{*,k} = V_0 Q_0 V_0^T$. Without loss of generality, if any diagonal element in $\Omega_{\infty,k}$ is non-zero, the corresponding element in $\mathbf{a}$ is set to zero. When $\kappa \to \infty$, the vector δ is said to be diffuse, which results in diffuse initialization of the Kalman filter.

2.5.1 Diffuse Priors

In general, diffuse initialization applies to the non-stationary components of the first order states, as the stationary components are well defined by model assumptions. In the context of time series, nothing is known about δ for the single starting point. Allowing infinite variance associated with the initial state reflects this lack of knowledge. The influence of this infinite variance is usually negligible after a small number of recursions, and the likelihood requires minor adjustments for numerical optimization.

In a stream network, first order reaches can enter anywhere in the network. This influences both the recursions as well as the likelihood. Similar to time series, the influence of the infinite variance may become negligible with progression downstream. However, the infinite variance may be re-introduced every time a lower order reach enters the network, greatly impacting the recursions. Furthermore, the infinite variance with every first order reach can easily dominate the likelihood, thus causing similar numerical problems in estimation. Consequently, the usefulness of a diffuse prior is a function of the network structure and the degree of non-stationary.

A common approach is to replace the scalar κ with an arbitrarily large number, but this numerical solution is not exact and may generate inaccuracies due to numerical rounding errors (Koopman, 1997, p.101). This also requires a modification to the likelihood in parameter estimation. The initialization approach of Durbin and Koopman (2001) uses a power series expansion in κ^{-1}, only keeping the first few terms, modifying the recursions and likelihood accordingly. We explore this technique adapted to the stream network and identify the complexities that may arise. Although the following recursions may be applicable for some models and network structures, it also seems reasonable to estimate the distribution of the initial state as an alternative to a diffuse prior since there are multiple occurrences of first order reaches. We present a method to assist in estimation of these initial

conditions using a concentrated likelihood in §3.2. Here we consider simple structures such as binary trees to demonstrate the applicability of a diffuse prior on a tree structure.

2.5.2 Exact Kalman Recursions

Durbin and Koopman (2001) refer to this modification as the "exact" Kalman filter. This approach is based on the variance decomposition

$$\Omega_k^p = \kappa \Omega_{\infty,k}^p + \Omega_{*,k}^p + O(\kappa^{-1}) \tag{2.26}$$

where $O(\kappa^{-j})$ denotes a matrix where each element is a function $f(\kappa)$ such that the limit of $\kappa^j f(\kappa)$ as $\kappa \longrightarrow \infty$ is finite.

Durbin and Koopman (2001, Section 4.2) derive the recursive relationship of filtering based on a single iterative step whereas we adopt that two step process of filtering by first predicting followed by an update. (Durbin and Koopman (2001, p.68) refer to this two step process as the so-called *contemporaneous filtering equations* and show their equivalence.) As they applied this decomposition to their single step filter, we apply this same decomposition to our two step filter.

Derivations for the following filter are analogous to those in Durbin and Koopman (2001, Section 5.2), where each step of the filtering process is re-formulated with the variance decomposition (2.26) and consequently a similar decomposition of Δ_k. The derivation of the smoother similarly follows, and uses the alternative formulation of the upstream recursions discussed in §2.4.3. Each of the following steps simply tracks powers of κ to identify terms that can be eliminated as κ tends to ∞. The formulations in Theorems (2.4.1) and (2.4.2) are addressed piecewise due to the number of required substitutions of these variance decompositions.

2.5.2.1 Exact Filter

To begin, we carry the decomposition in (2.26) through components in the initial prediction and updating equations. Since $\Delta_k = G_k \Omega_k^p G_k^T + R_k$ is a function of Ω_k^p, it too will have similar a similar decomposition defined by

$$\Delta_k = \kappa \Delta_{\infty,k} + \Delta_{*,k} + O(\kappa^{-1})$$

where

$$\Delta_{\infty,k} = G_k \Omega_{\infty,k}^p G_k^T, \qquad \Delta_{*,k} = G_k \Omega_{*,k}^p G_k^T + R_k$$

Two cases of this "exact" formulation are considered which depend on Δ_{∞,u_i}. We consider the cases where Δ is non-singular or zero as suggested by Durbin and Koopman (2001, p.102), where they justify this for a time series by three reasons. First, it provides a solution for the univariate case, since if $y(k)$ is univariate, then $\Delta_{\infty,k}$ must be positive or zero. Second, the restriction is satisfied in most practical situations when $\mathbf{y}(k)$ is multivariate. And lastly, in multivariate cases where this restriction is not satisfied, there are techniques such as treating the series as univariate that can be employed. The univariate case has direct application to a stream network, but this restriction for the multivariate case on a stream network has yet to be explored. We proceed as if this limitation is justifiable for the state-space model on a stream network.

The first case assumes that Δ_{∞,u_i} is nonsingular and is based on the expansion for $\Delta_{u_i}^{-1} = [\kappa \Delta_{\infty,u_i} + \Delta_{*,u_i} + O(\kappa^{-1})]^{-1}$ as a power series in κ^{-1}. This is

$$\Delta_{u_i}^{-1} = \Delta_{u_i}^{(0)} + \frac{1}{\kappa}\Delta_{u_i}^{(1)} + \frac{1}{\kappa^2}\Delta_{u_i}^{(2)} + O(\kappa^{-3})$$

for large κ where only $\Delta_{u_i}^{(j)}$ for $j = 0, 1, 2$ are needed in the reformulation of the recursions. The requirement of large κ is needed for the expansion to exist and can

be seen through univariate cases. For n-dimensional $\mathbf{Y}(u_i)$, $\Delta_{u_i}\Delta_{u_i}^{-1} = I_n$ results in

$$\begin{aligned} I_n &= \left(\kappa\Delta_{\infty,u_i} + \Delta_{*,u_i} + \kappa^{-1}\Delta_{u_i}^{(a)} + \kappa^{-2}\Delta_{u_i}^{(b)} + \cdots\right) \\ &\quad \times \left(\Delta_{u_i}^{(0)} + \kappa^{-1}\Delta_{u_i}^{(1)} + \kappa^{-2}\Delta_{u_i}^{(2)} + \cdots\right) \end{aligned}$$

where by matching coefficients of κ for $j = 1, 0, -1, -2$ we obtain

$$\begin{aligned} \Delta_{\infty,u_i}\Delta_{u_i}^{(0)} &= 0 \\ \Delta_{\infty,u_i}\Delta_{u_i}^{(1)} + \Delta_{*,u_i}\Delta_{u_i}^{(0)} &= I_n \\ \Delta_{\infty,u_i}\Delta_{u_i}^{(2)} + \Delta_{*,u_i}\Delta_{u_i}^{(1)} + \Delta_{u_i}^{(a)}\Delta_{u_i}^{(0)} &= 0 \\ \Delta_{*,u_i}\Delta_{u_i}^{(2)} + \Delta_{u_i}^{(a)}\Delta_{u_i}^{(1)} + \Delta_{u_i}^{(b)}\Delta_{u_i}^{(0)} &= 0, \text{ etc.} \end{aligned}$$

To solve these equations for $\Delta_{u_i}^{(j)}$, $j = 0, 1, 2$, we only consider the case where Δ_{∞,u_i} is nonsingular. Since Δ_{∞,u_i} is nonsingular, it must be that $\Delta_{u_i}^{(0)} = 0$ by the first equation. Simplification of the remaining equations results in

$$\Delta_{u_i}^{(0)} = 0 \qquad \Delta_{u_i}^{(1)} = \Delta_{\infty,u_i}^{-1} \qquad \Delta_{u_i}^{(2)} = -\Delta_{\infty,u_i}^{-1}\Delta_{*,u_i}\Delta_{\infty,u_i}^{-1}.$$

In the case when $\Delta_{\infty,u_i} = 0$, we see that

$$\Delta_{u_i}^{-1} = \Delta_{*,u_i}^{-1} + O(\kappa^{-1}).$$

From (2.10), there are a number of manipulations required since $\Omega_{u_i}^p$ and Δ_{u_i} are reformulated to terms in powers of κ. Expressing $\mathcal{H}(k, u_i) = F_{k,u_i}\Omega_{u_i}^p G_{u_i}^T$ with Ω_k^p in the form of (2.26), when Δ_{∞,u_i} is positive definite we have

$$\begin{aligned} \mathcal{H}(k, u_i) &= F_{k,u_i}\left[\kappa\Omega_{\infty,u_i}^p + \Omega_{*,u_i}^p + O(\kappa^{-1})\right] G_{u_i}^T \\ &= \kappa F_{k,u_i}\Omega_{\infty,u_i}^p G_{u_i}^T + F_{k,u_i}\Omega_{*,u_i}^p G_{u_i}^T + O(\kappa^{-1}) \\ &= \kappa\mathcal{H}_\infty(k, u_i) + \mathcal{H}_*(k, u_i) + O(\kappa^{-1}) \end{aligned}$$

from which we obtain

$$\mathcal{H}(k, u_i)\Delta_{u_i}^{-1} = \left(\kappa\mathcal{H}_\infty(k, u_i) + \mathcal{H}_*(k, u_i) + O(\kappa^{-1})\right) \times \left(\kappa^{-1}\Delta_{u_i}^{(1)} + \kappa^{-2}\Delta_{u_i}^{(2)} + O(\kappa^{-3})\right)$$

$$= \mathcal{H}_\infty(k,u_i)\Delta^{(1)}_{u_i} + \kappa^{-1}\left(\mathcal{H}_\infty(k,u_i)\Delta^{(2)}_{u_i} + \mathcal{H}_*(k,u_i)\Delta^{(1)}_{u_i}\right) + O(\kappa^{-2}). \quad (2.27)$$

When $\Delta_{\infty,u_i} = 0$, we have

$$\mathcal{H}(k,u_i) = F_{k,u_i}\Omega^p_{*,u_i}G^T_{u_i} + O(\kappa^{-1})$$

and

$$\mathcal{H}(k,u_i)\Delta^{-1}_{u_i} = F_{k,u_i}\Omega^p_{*,u_i}G^T_{u_i}\Delta^{-1}_{*,u_i} + O(\kappa^{-1}) \quad (2.28)$$

since it must be that $\Omega^p_{\infty,u_i}G^T_{u_i} = 0$.

Following Durbin and Koopman (2001, p.103), we see from the recursions in Theorem 2.4.1 that $\mathbf{X}^p(k)$ has the form

$$\mathbf{X}^p(k) = \mathbf{X}^p(k)^{(0)} + \kappa^{-1}\mathbf{X}^p(k)^{(1)} + O(\kappa^{-2}) \quad (2.29)$$

where $\mathbf{X}^p(k)^{(0)} = \mathbf{a}$ and $\mathbf{X}^p(k)^{(1)} = 0$ for any first order reach k. Consequently, it is easy to see that the innovation

$$\begin{aligned}\mathbf{v}(k) &= \mathbf{Y}(k) - G_k\left(\mathbf{X}^p(k)^{(0)} + \kappa^{-1}\mathbf{X}^p(k)^{(1)} + O(\kappa^{-2})\right)\\ &= \mathbf{Y}(k) - G_k\mathbf{X}^p(k)^{(0)} + \kappa^{-1}\left(-G_k\mathbf{X}^p(k)^{(1)}\right) + O(\kappa^{-2})\\ &= \mathbf{v}(k)^{(0)} + \kappa^{-1}\mathbf{v}(k)^{(1)} + O(\kappa^{-2}). \quad (2.30)\end{aligned}$$

(Any reference to $\mathbf{v}(k)$ immediately applies to the innovation and should not be confused with the unobservable state noise $\mathbf{V}(k)$.)

For parent reach u_i with Δ_{∞,u_i} positive definite, we see that

$$\begin{aligned}\mathcal{H}(k,u_i)\Delta^{-1}_{u_i}\mathbf{v}(u_i) &= \left(\mathcal{H}_\infty(k,u_i)\Delta^{(1)}_{u_i} + \kappa^{-1}\left(\mathcal{H}_\infty(k,u_i)\Delta^{(2)}_{u_i} + \mathcal{H}_*(k,u_i)\Delta^{(1)}_{u_i}\right) + O(\kappa^{-2})\right)\\ &\quad \times \left(\mathbf{v}(u_i)^{(0)} + \kappa^{-1}\mathbf{v}(u_i)^{(1)} + O(\kappa^{-2})\right)\\ &= \mathcal{H}_\infty(k,u_i)\Delta^{(1)}_{u_i}\mathbf{v}(u_i)^{(0)}\\ &\quad + \kappa^{-1}\left(\mathcal{H}_\infty(k,u_i)\Delta^{(2)}_{u_i} + \mathcal{H}_*(k,u_i)\Delta^{(1)}_{u_i}\right)\mathbf{v}(u_i)^{(0)}\end{aligned}$$

$$+\kappa^{-1}\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(1)}\mathbf{v}(u_i)^{(1)}+O(\kappa^{-2})$$

where

$$\mathcal{H}(k,u_i)\Delta_{u_i}^{-1}\mathbf{v}(u_i)=F_{k,u_i}\Omega_{*,u_i}^{p}G_{u_i}^{T}\Delta_{\star,u_i}^{-1}\mathbf{v}^{(0)}(u_i)+O(\kappa^{-1})$$

when $\Delta_{\infty,u_i}=0$. From these we see that $F_{k,u_i}\mathbf{X}^{p}(u_i)+\mathcal{H}(k,u_i)\Delta_{u_i}^{-1}\mathbf{v}(u_i)$ becomes

$$\begin{cases} F_{k,u_i}\mathbf{X}^{p}(u_i)^{(0)}+\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(1)}\mathbf{v}(u_i)^{(0)}, & \Delta_{\infty,u_i}\text{ positive definite} \\ F_{k,u_i}\mathbf{X}^{p}(u_i)^{(0)}+F_{k,u_i}\Omega_{*,u_i}^{p}G_{u_i}^{T}\Delta_{\star,u_i}^{-1}\mathbf{v}^{(0)}(u_i), & \Delta_{\infty,u_i}=0 \end{cases} \tag{2.31}$$

as $\kappa\to\infty$. Then by (2.10), $\mathbf{X}^{p}(k)^{(0)}$ can be determined accordingly by summing (2.31) for each parent to obtain a prediction similar in form to the filtered prediction in Durbin and Koopman (2001, p.103).

The recursive representation for $\Omega_{\infty,k}^{p}$ and $\Omega_{*,k}^{p}$ is derived from (2.11). Using (2.27), we find for Δ_{∞,u_i} positive definite that

$$\begin{aligned} \mathcal{H}(k,u_i)\Delta_{u_i}^{-1}\mathcal{H}(k,u_i)^{T} &= \left(\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(1)}+\kappa^{-1}\left(\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(2)}+\mathcal{H}_{*}(k,u_i)\Delta_{u_i}^{(1)}\right)+O(\kappa^{-2})\right) \\ &\quad\times\left(\kappa\mathcal{H}_{\infty}(k,u_i)^{T}+\mathcal{H}_{*}(k,u_i)^{T}+O(\kappa^{-1})\right) \\ &= \kappa\left(\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(1)}\mathcal{H}_{\infty}(k,u_i)^{T}\right)+\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(2)}\mathcal{H}_{\infty}(k,u_i)^{T} \\ &\quad+\mathcal{H}_{*}(k,u_i)\Delta_{u_i}^{(1)}\mathcal{H}_{\infty}(k,u_i)^{T}+\mathcal{H}_{\infty}(k,u_i)\Delta_{u_i}^{(1)}\mathcal{H}_{*}(k,u_i)^{T} \\ &\quad+O(\kappa^{-1}) \end{aligned} \tag{2.32}$$

whereas with (2.28) we see that

$$\mathcal{H}(k,u_i)\Delta_{u_i}^{-1}\mathcal{H}(k,u_i)^{T}=F_{k,u_i}\Omega_{*,u_i}^{p}G_{u_i}^{T}\Delta_{\star,u_i}^{-1}G_{u_i}\Omega_{*,u_i}^{p}F_{k,u_i}^{T}+O(\kappa^{-1}) \tag{2.33}$$

when $\Delta_{\infty,u_i}=0$. With that, appropriate substitution in

$$F_{k,u_i}\Omega_{u_i}^{p}F_{k,u_i}^{T}-\mathcal{H}(k,u_i)\Delta_{u_i}^{-1}\mathcal{H}(k,u_i)^{T}$$

and summing over each parent will lead to the recursive relationship in $\Omega_{\infty,k}^{p}$ and $\Omega_{*,k}^{p}$ since the contribution of parent u_i will depend on Δ_{∞,u_i}. There are three possibilities, either both Δ_{∞,u_i} are positive definite, both zero, or one zero while the

other is positive definite. In any of these cases, (2.32) and (2.33) can be used to obtain the necessary prediction variance decomposition.

Similar arguments are used to derive the exact recursions for the updated estimates and variances. Needed in (2.12) and (2.13), we find for $\Delta_{\infty,k}$ positive definite that

$$\begin{aligned}\Omega_k^p G_k^T \Delta_k^{-1} &= \left(\kappa\Omega_{\infty,k}^p G_k^T + \Omega_{*,k}^p G_k^T + O(\kappa^{-1})\right) \times \left(\kappa^{-1}\Delta_k^{(1)} + \kappa^{-2}\Delta_k^{(2)} + O(\kappa^{-3})\right) \\ &= \Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} + \kappa^{-1}\left(\Omega_{\infty,k}^p G_k^T \Delta_k^{(2)} + \Omega_{*,k}^p G_k^T \Delta_k^{(1)}\right) + O(\kappa^{-2})\end{aligned}$$

and

$$\begin{aligned}\Omega_k^p G_k^T \Delta_k^{-1} G_k \Omega_k^p &= \left(\Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} + \kappa^{-1}\left(\Omega_{\infty,k}^p G_k^T \Delta_k^{(2)} + \Omega_{*,k}^p G_k^T \Delta_k^{(1)}\right) + O(\kappa^{-2})\right) \\ &\quad \times \left(\kappa G_k \Omega_{\infty,k}^p + G_k \Omega_{*,k}^p + O(\kappa^{-1}\right) \\ &= \kappa\Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p + \Omega_{\infty,k} G_k^T \Delta_k^{(1)} G_k \Omega_{*,k}^p + \Omega_{\infty,k}^p G_k^T \Delta_k^{(2)} G_k \Omega_{\infty,k}^p \\ &\quad + \Omega_{*,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p + O(\kappa^{-1}).\end{aligned}$$

From these we find

$$\mathbf{X}^f(k)^{(0)} = \mathbf{X}^p(k)^{(0)} + \Omega_{\infty,k} G_k^T \Delta_k^{(1)} \mathbf{v}(k)^{(0)}$$

with variance $\Omega_k^f = \kappa\Omega_{\infty,k}^f + \Omega_{*,k}^f + O(\kappa^{-1})$ since

$$\begin{aligned}\Omega_k^f &= \Omega_k^p - \Omega_k^p G_k^T \Delta_k^{-1} G_k \Omega_k^p \\ &= \kappa\Omega_{\infty,k}^p + \Omega_{*,k}^p - \kappa\Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p - \Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{*,k}^p \\ &\quad - \Omega_{\infty,k}^p G_k^T \Delta_k^{(2)} G_k \Omega_{\infty,k}^p - \Omega_{*,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p + O(\kappa^{-1}) \\ &= \kappa\left(\Omega_{\infty,k}^p - \Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p\right) + \Omega_{*,k}^p \\ &\quad - \Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{*,k}^p - \Omega_{\infty,k}^p G_k^T \Delta_k^{(2)} G_k \Omega_{\infty,k}^p - \Omega_{*,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p + O(\kappa^{-1}).\end{aligned}$$

The recursive decomposition of the filtering variance becomes

$$\begin{aligned}\Omega_{\infty,k}^f &= \Omega_{\infty,k}^p - \Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p \\ \Omega_{*,k}^f &= \Omega_{*,k}^p - \Omega_{\infty,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{*,k}^p - \Omega_{\infty,k}^p G_k^T \Delta_k^{(2)} G_k \Omega_{\infty,k}^p - \Omega_{*,k}^p G_k^T \Delta_k^{(1)} G_k \Omega_{\infty,k}^p\end{aligned}$$

as $\kappa \to \infty$. When $\Delta_{\infty,k} = 0$, there is much simplification, and we see that

$$\Omega_k^p G_k^T \Delta_k^{-1} = \Omega_{*,k}^p G_k^T \Delta_{*,k}^{-1} + O(\kappa^{-1})$$

and

$$\Omega_k^p G_k^T \Delta_k^{-1} G_k \Omega_k^p = \Omega_{*,k}^p G_k^T \Delta_{*,k}^{-1} G_k \Omega_{*,k}^p + O(\kappa^{-1}).$$

from which we see a simplification in updating the prediction variances. The resulting updates are

$$\begin{aligned} \Omega_{\infty,k}^f &= \Omega_{\infty,k}^p \\ \Omega_{*,k}^f &= \Omega_{*,k}^p - \Omega_{*,k}^p G_k^T \Delta_{*,k}^{-1} G_k^p \Omega_{*,k} + Q_k \end{aligned}$$

as $\kappa \to \infty$ with $\Delta_{\infty,k} = 0$.

In either case for Δ_{∞,u_i} and $\Delta_{\infty,k}$, the initial predictions can now be expressed in terms of previous filtered values

$$\mathbf{X}^p(k)^{(0)} = F_{k,u_1} \mathbf{X}^f(u_1)^{(0)} + F_{k,u_2} \mathbf{X}^f(u_2)^{(0)} \tag{2.34}$$

with variance decomposition

$$\begin{aligned} \Omega_{\infty,k}^p &= F_{k,u_1} \Omega_{\infty,u_1}^f F_{k,u_1}^T + F_{k,u_2} \Omega_{\infty,u_2}^f F_{k,u_2}^T \\ \Omega_{*,k}^p &= F_{k,u_1} \Omega_{*,u_1}^f F_{k,u_1}^T + F_{k,u_2} \Omega_{*,u_2}^f F_{k,u_2}^T + Q_k \end{aligned} \tag{2.35}$$

both of which are analogous to those derived in Durbin and Koopman (2001, Section 5.2) with the necessary modifications due to the tree structure of a stream network.

2.5.2.2 Transition to the Usual Kalman Filter

As seen in the above derivations, the variance decomposition influences each step as the recursions progress downstream. Durbin and Koopman (2001, Section 5.2.2) argue that the conditional variance of $\delta|\mathbf{Y}$ should be finite, suggesting there exists some point in the recursions such that all future predictions will have finite variances, given the previous observations. Application of this argument with stream networks is much more complex than that in time series. As we will see, the applicability of a diffuse prior on a stream network is primarily a function of the physical structure and nonstationary model assumed.

We suppose a simple tree structure, one in which lower order reaches eventually do not re-enter the network, such as a binary tree. More specifically, we consider a tree in which we can identify b disjoint initial basins $\mathcal{B}_j, j = 1 \ldots b$ defined in such a way that guarantees finite Ω_k^p for every k downstream of each initial basin yet has infinite components for some reaches within these initial basins. The premise is that filtering and smoothing estimates within each initial basin need adjustment because of the components that tend to infinity with κ. Outside of these initial basins, no adjustment is needed since all components of Ω_k^p are finite and independent of κ.

Basins are defined in the following manner. In the simplest of nonstationary models with δ a scalar, a basin can be a single first order reach. As long as there exists an observation, the filtering variance will be finite. For more complicated models, a basin will be defined by a set of reaches such that there exists two furthest downstream reaches, both with finite filtering variance, that merge. In either case, there exists a downstream reach, d_j, which flows directly out of the basin such that $\Omega_{d_j}^p < \infty$. Defining basins in this manner guarantees disjoint sets of reaches, and that $\Omega_k^p < \infty$ for all k downstream of any d_j. The size of each $\mathcal{B}_j$ is a function of the physical tree structure and the nonstationary model of interest. Furthermore,

the existence of missing data greatly influences the ability to define disjoint basins needed for use of a diffuse prior.

Finite $\text{Var}[\delta|\mathcal{B}_{ij}]$ is still used in an argument for finite smoothed predictions and variances. In §3.2, an expression for $\log p(\delta, \mathbf{Y})$ is derived that is independent of the initial states and can be used to show the existence of $\text{Var}[\delta|\mathbf{y}(k), k \in \mathcal{B}_{ij}]$. Since $p(\delta|\mathbf{Y}(k), k \in \mathcal{B}_{ij})$ is Gaussian, $\log p(\delta|\mathbf{y}(k), k \in \mathcal{B}_{ij})$ is quadratic in δ so its second derivative does not depend on δ. By definition, the reciprocal of the second derivative of $\log p(\delta|\mathbf{y}(k), k \in \mathcal{B}_{ij})$ with respect to δ exists, implying finite conditional variance.

Because $\Omega_k^p < \infty$ for $k \geq d_j, j = 1...b$, $\Omega_{\infty,k}^p = 0$, all predictions downstream of any d_j will have finite prediction error. With that, the usual Kalman filter can be used, setting $\Omega_{d_j}^p = \Omega_{\star,d_j}^p$.

2.5.2.3 Exact Kalman Smoother

The backwards recursions developed in §2.4.3 are used in formulating the exact smoother on a stream network. Let d_j identify the exiting reach in each local basin such that $\Omega_{\infty,k}^p = 0$ for all $k \geq d_j$. This implies that the only smoothed estimates influenced by κ are those that are upstream of d_j. Therefore the formulations for the smoothed estimates and variances for $k < d_j, j = 1...b$ are presented whereas the smoothed estimates for all other k are obtained via the usual Kalman smoother. The following formulations are analogous to those provided in Durbin and Koopman (2001, Section 5.3) with appropriate modification due to the tree structure of a stream network.

Using the power series expansion for $\Delta_{u_i}^{-1}$ when Δ_{∞,u_i} is positive definite, it is easy to show that $L_{k,u_i} = L_{k,u_i}^{(0)} + \kappa^{-1} L_{k,u_i}^{(1)} + O(\kappa^{-2})$ where

$$\begin{aligned} L_{k,u_i}^{(0)} &= F_{k,u_i} - F_{k,u_i}\Omega_{\infty,u_i}^p G_{u_i}^T \Delta_{\infty,u_i}^{-1} G_{u_i} \\ L_{k,u_i}^{(1)} &= F_{k,u_i}\Omega_{\infty,u_i}^p G_{u_i}^T \Delta_{\infty,u_i}^{-1} \Delta_{\star,u_i} \Delta_{\infty,u_i}^{-1} G_{u_i} - F_{k,u_i}\Omega_{\star,u_i}^p G_k^T \Delta_{\infty,u_i}^{-1} G_{u_i}. \end{aligned}$$

Since the usual recursions are used for $k \geq d_j$, define $\mathbf{r}^{(0)}(d_j) = \mathbf{r}(d_j)$ and $\mathbf{r}^{(1)}(d_j) = 0$, and consider expressing the recursions upstream of d_j in terms of power series expansions in κ^{-1}. Since $\mathbf{r}(u_i) = G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) + L_{k,u_i}^T \mathbf{r}(k)$, it is easy to verify that

$$\mathbf{r}(u_i) = \mathbf{r}^{(0)}(u_i) + \kappa^{-1}\mathbf{r}^{(1)}(u_i) + O(\kappa^{-2}), \text{ for } k < d_j, j = 1, \ldots, b$$

where

$$\begin{aligned} \mathbf{r}^{(0)}(u_i) &= L_{k,u_i}^{(0)T} \mathbf{r}^{(0)}(k) \\ \mathbf{r}^{(1)}(u_i) &= G_{u_i}^T \Delta_{\infty,u_i}^{-1} \mathbf{v}^{(0)}(u_i) + L_{k,u_i}^{(0)T} \mathbf{r}^{(1)}(k) + L_{k,u_i}^{(1)T} \mathbf{r}^{(0)}(k) \end{aligned}$$

which is analogous and similar in form to Durbin and Koopman (2001, Eq (5.21)). Using (2.22), the smoothed state vector is

$$\begin{aligned} \mathbf{X}^s(u_i) &= \mathbf{X}^p(u_i) + \Omega_{u_i}^p \mathbf{r}(u_i) \\ &= \mathbf{X}^p(u_i) + \left(\kappa\Omega_{\infty,u_i}^p + \Omega_{\star,u_i}^p + O(\kappa^{-1})\right) \times \left(\mathbf{r}^{(0)}(u_i) + \kappa^{-1}\mathbf{r}^{(1)}(u_i) + O(\kappa^{-2})\right) \\ &= \mathbf{X}^p(u_i) + \kappa\Omega_{\infty,u_i}^p \mathbf{r}^{(0)}(u_i) + \Omega_{\infty,u_i}^p \mathbf{r}^{(1)}(u_i) + \Omega_{\star,u_i}^p \mathbf{r}^{(0)}(u_i) + O(\kappa^{-1}) \end{aligned}$$

where it is obvious that it must be that $\Omega_{\infty,u_i}^p \mathbf{r}^{(0)}(u_i) = 0$ for every reach for this to make sense. Using the arguments of Durbin and Koopman (2001, p.106), it is clear that $\Omega_{\infty,u_i}^p \mathbf{r}^{(0)}(u_i) \neq 0$ guarantees $\text{Var}(\mathbf{X}(u_i)|\mathbf{Y}) \to \infty$ as $\kappa \to \infty$. This implies that $\Omega_{\infty,u_i}^p \mathbf{r}^{(0)}(u_i) = 0$ at every reach is necessary for finite conditional variance. Thus, it suffices to show that if $\text{Var}(\mathbf{X}(u_i)|\mathbf{Y}) < \infty$, then $\Omega_{\infty,u_i}^p \mathbf{r}^{(0)}(u_i) = 0$. Now, variation in $X(u_i)$ comes from δ and the additional noise components. Since the finite number of noise components each have finite variances, they will also have finite conditional variances. Furthermore, by the existence of a reach d_j such that $\text{Var}\big(\delta|\mathbf{Y}(k'), k' \in \mathcal{U}_{(d_j)}\big) < \infty$, then $\text{Var}(\delta|\mathbf{Y}) < \infty$. Hence, it must be the case that $\text{Var}(\mathbf{X}(u_i)|\mathbf{Y}) < \infty$. Letting $\kappa \to \infty$, we then have

$$\mathbf{X}^s(u_i) = \mathbf{X}^p(u_i)^{(0)} + \Omega_{\infty,u_i}^p \mathbf{r}^{(1)}(u_i) + \Omega_{\star,u_i}^p \mathbf{r}^{(0)}(u_i), \text{ for } u_i < d_j, j = 1, \ldots, b \quad (2.36)$$

with $\mathbf{r}^{(0)}(d_j) = \mathbf{r}(d_j)$ and $\mathbf{r}^{(1)}(d_j) = 0$ since the downstream recursions do not depend on κ.

Following (Durbin and Koopman, 2001, Section 5.3.2), we see from §2.4.3 that $\Omega^s_{u_i} = \Omega^p_{u_i} - \Omega^p_{u_i} N_{u_i} \Omega^p_{u_i}$ with $N_{u_i} = G^T_{u_i} \Delta^{-1}_{u_i} G_{u_i} + L^T_{k,u_i} N_k L_{k,u_i}$. As with the smoothed predictions for reaches upstream of d_j, the corresponding prediction variances also need to be modified when using the diffuse prior on δ. Writing N_{u_i} as a power series expansion in κ^{-1}, we have

$$N_{u_i} = N^{(0)}_{u_i} + \kappa^{-1} N^{(1)}_{u_i} + \kappa^{-2} N^{(2)}_{u_i} + O(\kappa^{-3})$$

which can be used recursively in

$$\begin{aligned} N_{u_i} &= G^T_{u_i}(\kappa^{-1}\Delta^{(1)}_{u_i} + \kappa^{-2}\Delta^{(2)}_{u_i}) G_{u_i} + \left(L^{(0)}_{k,u_i} + \kappa^{-1} L^{(1)}_{k,u_i} + \kappa^{-2} L^{(2)}_{k,u_i}\right)^T \\ &\quad \times \left(N^{(0)}_k + \kappa^{-1} N^{(1)}_k + \kappa^{-2} N^{(2)}_k\right) \left(L^{(0)}_{k,u_i} + \kappa^{-1} L^{1)}_{k,u_i} + \kappa^{-2} L^{(2)}_{k,u_i}\right) \end{aligned}$$

when the residual terms are ignored. From this we see

$$\begin{aligned} N^{(0)}_{u_i} &= L^{(0)T}_{k,u_i} N^{(0)}_k L^{(0)}_{k,u_i} \\ N^{(1)}_{u_i} &= G^T_{u_i} \Delta^{(1)}_{u_i} G_{u_i} + L^{(0)T}_{k,u_i} N^{(1)}_k L^{(0)}_{k,u_i} + L^{(1)T}_{k,u_i} N^{(0)}_k L^{(0)}_{k,u_i} + L^{(0)T}_{k,u_i} N^{(0)}_k L^{(1)}_{k,u_i} \\ N^{(2)}_{u_i} &= G^T_{u_i} \Delta^{(2)}_{u_i} G_{u_i} + L^{(0)T}_{k,u_i} N^{(2)}_k L^{(0)}_{k,u_i} + L^{(1)T}_{k,u_i} N^{(1)}_k L^{(0)}_{k,u_i} + L^{(2)T}_{k,u_i} N^{(0)}_k L^{(0)}_{k,u_i} \\ &\quad L^{(0)T}_{k,u_i} N^{(1)}_k L^{(1)}_{k,u_i} + L^{(1)T}_{k,u_i} N^{(0)}_k L^{(1)}_{k,u_i} + L^{(0)T}_{k,u_i} N^{(0)}_k L^{(2)}_{k,u_i} \end{aligned}$$

where $N^{(0)}_{d_j} = N_{d_j}$ and $N^{(1)}_{d_j} = N^{(2)}_{d_j} = 0$ since recursions downstream of d_j do not depend on κ. Notice again that these equations are analogous to (Durbin and Koopman, 2001, Eq. 5.26). With that and again ignoring the residual terms, we have

$$\begin{aligned} \Omega^s_{u_i} &= \Omega^p_{u_i} - \Omega^p_{u_i} N_{u_i} \Omega^p_{u_i} \\ &= \kappa \Omega^p_{\infty,u_i} + \Omega^p_{\star,u_i} \\ &\quad - \left(\kappa \Omega^p_{\infty,u_i} + \Omega^p_{\star,u_i}\right) \times \left(N^{(0)}_{u_i} + \kappa^{-1} N^{(1)}_{u_i} + \kappa^{-2} N^{(2)}_{u_i}\right) \times \left(\kappa \Omega^p_{\infty,u_i} + \Omega^p_{\star,u_i}\right) \\ &= -\kappa^2 \Omega^p_{\infty,u_i} N^{(0)}_{u_i} \Omega^p_{\infty,u_i} \end{aligned}$$

$$
\begin{aligned}
&+\kappa\left(\Omega^p_{\infty,u_i} - \Omega^p_{\infty,u_i}N^{(0)}_{u_i}\Omega^p_{\star,u_i} - \Omega^p_{\star,u_i}N^{(0)}_{u_i}\Omega^p_{\infty,u_i} - \Omega^p_{\infty,u_i}N^{(1)}_{u_i}\Omega^p_{\infty,u_i}\right)\\
&+\Omega^p_{\star,u_i} - \Omega^p_{\star,u_i}N^{(0)}_{u_i}\Omega^p_{\star,u_i} - \Omega^p_{\star,u_i}N^{(1)}_{u_i}\Omega^p_{\infty,u_i} - \Omega^p_{\infty,u_i}N^{(1)}_{u_i}\Omega^p_{\star,u_i}\\
&-\Omega^p_{\infty,u_i}N^{(2)}_{u_i}\Omega^p_{\infty,u_i}.
\end{aligned}
$$

Using the same argument about the κ term in the exact smoother, we see that the matrix terms with coefficients κ and κ^2 must be zero if $\mathrm{Var}[\mathbf{X}(k)|\mathbf{Y}] < \infty$. When $\kappa \to \infty$, the smoothed state variance becomes

$$
\begin{aligned}
\Omega^s_{u_i} &= \Omega^p_{\star,u_i} - \Omega^p_{\star,u_i}N^{(0)}_{u_i}\Omega^p_{\star,u_i} - \Omega^p_{\star,u_i}N^{(1)}_{u_i}\Omega^p_{\infty,u_i}\\
&\quad -\Omega^p_{\infty,u_i}N^{(1)}_{u_i}\Omega^p_{\star,u_i} - \Omega^p_{\infty,u_i}N^{(2)}_{u_i}\Omega^p_{\infty,u_i} \qquad (2.37)
\end{aligned}
$$

when reach u_i is upstream of d_j.

For the case when $\Delta_{\infty,u_i} = 0$, we consider one more term in the expansions

$$
\begin{aligned}
\Omega^p_{u_i} &= \kappa\Omega^p_{\infty,u_i} + \Omega^p_{\star,u_i} + \kappa^{-1}\Omega^p_{a,u_i} + O(\kappa^{-2})\\
\Delta^{-1}_{u_i} &= \Delta^{-1}_{\star,u_i} + \kappa^{-1}\Delta^{(a)}_{u_i} + O(\kappa^{-2})
\end{aligned}
$$

as suggested in Koopman and Durbin (2003, Appendix II, derivation of equation (31)). Using this form of the expansion, we see that

$$
\begin{aligned}
L^{(0)}_{k,u_i} &= F_{k,u_i} - F_{k,u_i}\Omega^p_{\star,u_i}G^T_{u_i}\Delta^{-1}_{\star,u_i}G_{u_i}\\
L^{(1)}_{k,u_i} &= -\left(F_{k,u_i}\Omega^p_{a,u_i}G^T_{u_i}\Delta^{-1}_{\star,u_i}G_{u_i} + F_{k,u_i}\Omega^p_{\star,u_i}G^T_{u_i}\Delta^{(a)}_{\star,u_i}G_{u_i}\right)
\end{aligned}
$$

where we notice the leading term in $L^{(1)T}_{k,u_i}$ is $G^T_{u_i}$. Substitution in $\mathbf{r}(u_i)$ leads to

$$
\begin{aligned}
\mathbf{r}^{(0)}(u_i) &= G^T_{u_i}\Delta^{-1}_{\star,u_i}\mathbf{v}^{(0)}(u_i) + L^{(0)T}_{k,u_i}\mathbf{r}^{(0)}(k)\\
\mathbf{r}^{(1)}(u_i) &= G^T_{u_i}\Delta^{-1}_{\star,u_i}\mathbf{v}^{(1)}(u_i) + G^T_{u_i}\Delta^{(a)}_{\star,u_i}\mathbf{v}^{(0)}(u_i) + L^{(1)T}_{k,u_i}\mathbf{r}^{(0)}(k) + L^{(0)T}_{k,u_i}\mathbf{r}^{(1)}(k)
\end{aligned}
$$

Since $\mathbf{r}^{(1)}(u_i)$ is premultiplied by Ω^p_{∞,u_i} in (2.36), the smoothed estimate can be simplified to

$$
\mathbf{X}^s(u_i) = \mathbf{X}^p(u_i)^{(0)} + \Omega^p_{\star,u_i}\mathbf{r}^{(0)}(u_i) + \Omega^p_{\infty,u_i}L^{(0)T}_{k,u_i}\mathbf{r}^{(1)}(k)
$$

when $\Delta_{\infty,u_i} = 0$. Furthermore, substitution of alternative expansions for $\Omega^p_{u_i}$ and $\Delta^{-1}_{u_i}$ results in

$$\begin{aligned}
N^{(0)}_{u_i} &= G^T_{u_i}\Delta^{-1}_{u_i}G_{u_i} + L^{(0)T}_{k,u_i}N^{(0)}_k L^{(0)}_{k,u_i} \\
N^{(1)}_{u_i} &= G^T_{u_i}\Delta^{(a)}_{\star,u_i}G_{u_i} + L^{(0)T}_{k,u_i}N^{(1)}_k L^{(0)}_{k,u_i} + L^{(1)T}_{k,u_i}N^{(0)}_k L^{(0)}_{k,u_i} + L^{(0)T}_{k,u_i}N^{(0)}_k L^{(1)}_{k,u_i} \\
N^{(2)}_{u_i} &= G^T_{u_i}\Delta^{(b)}_{\star,u_i}G_{u_i} + L^{(0)T}_{k,u_i}N^{(2)}_k L^{(0)}_{k,u_i} + L^{(1)T}_{k,u_i}N^{(1)}_k L^{(0)}_{k,u_i} + L^{(2)T}_{k,u_i}N^{(0)}_k L^{(0)}_{k,u_i} \\
&\quad L^{(0)T}_{k,u_i}N^{(1)}_k L^{(1)}_{k,u_i} + L^{(1)T}_{k,u_i}N^{(0)}_k L^{(1)}_{k,u_i} + L^{(0)T}_{k,u_i}N^{(0)}_k L^{(2)}_{k,u_i}
\end{aligned}$$

which can be used in (2.37) to determine the variance of the smoothed estimate. Here we see that these equations are analogous to Koopman and Durbin (2003, Eq. 36). Simplification of this variance is possible since $G^T_{u_i}$ is the leading term in $L^{(1)}_{k,u_i}$ and $L^{(2)}_{k,u_i}$, which when premultiplied by Ω^p_{∞,u_i} eliminates that contribution since $\Omega^p_{\infty,u_i}G^T_{u_i} = 0$.

Remark 2.5.1 Another option to consider may be to recursively update parameter estimates for the initial state model. By this, we could use information from one basin to begin recursions in another basin, in hopes of maintaining a finite prediction error variance. However, identification of which basin to begin conditioning is unclear, so this option is not considered further here.

Chapter 3

THE GAUSSIAN LIKELIHOOD

The state-space representation adopted for the stream-network data allows for expression of the Gaussian likelihood function in terms of the orthogonal innovations and corresponding prediction error variances. To accomplish this, the logical order of flow must be determined. Define $\mathcal{S}_i$ to be the set of all i^{th} order stream *segments*. Since segments are disjoint, and each segment is ordered by flow, the sets $\mathcal{S}_i$ are a medium over which the Kalman recursions can be executed. Recalling that m_s is the highest Strahler order of all reaches in the network and that $n_{i,j}$ is defined to be the number of reaches on segment j of order i, the log of the Gaussian likelihood can be expressed as

$$l(\mathbf{y}|\theta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{m_s}\sum_{j\in\mathcal{S}_i}\sum_{k=1}^{n_{ij}}\left(\log|\Delta_k| + \mathbf{v}(k)^T\Delta_k^{-1}\mathbf{v}(k)\right) \qquad (3.1)$$

where $\mathbf{y}$ is the vector of observations, N is the total number of observations, $\mathbf{v}(k) = \mathbf{y}(k) - G_k\mathbf{X}^p(k)$, and $\Delta_k = G_k\Omega_k^pG_k^T + R_k$. These innovations and variances are obtained from execution of the Kalman Filter over the entire network.

Even in cases of non-Gaussian error structures, we still refer to (3.1) as the "Gaussian likelihood", as did Brockwell and Davis (1991, p.255) where they indicate that the Gaussian likelihood still serves as a measure of the goodness of fit of the covariance structure to the data. Therefore, it is still a reasonable criterion function to maximize in order to obtain parameter estimates.

3.1 Diffuse Likelihood

In the case where a diffuse initialization is considered, the log likelihood will not converge as $\kappa \to \infty$. Durbin and Koopman (2001) present a method of obtaining a *diffuse likelihood* for a time series that is adapted here to the tree structure of the stream network. Construction of the diffuse likelihood is subject to the same constraints as those assumed for utilizing a diffuse prior, mainly, an appropriate nonstationary model for the specific network structure.

The goal of a diffuse likelihood is to obtain a function $\log L_d(\mathbf{y}|\theta)$ that can be used rather than $\log L(\mathbf{y}|\theta)$ to obtain parameter estimates. In order to offset inflation from $\kappa \to \infty$, terms that are independent of the parameter vector are added to the log of the likelihood function.

To see this, express (3.1) as sums over disjoint sets by

$$\begin{aligned} l(\mathbf{Y}|\theta) &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{k\in\cup\mathcal{B}_j}\left(\log|\Delta_k| + \mathbf{v}(k)^T\Delta_k^{-1}\mathbf{v}(k)\right) \\ &\quad -\frac{1}{2}\sum_{k\notin\cup\mathcal{B}_j}\left(\log|\Delta_k| + \mathbf{v}(k)^T\Delta_k^{-1}\mathbf{v}(k)\right) \end{aligned}$$

from which it is clear that $l(\mathbf{Y}|\theta)$ diverges as $\kappa \to \infty$ since $|\Delta_k| \to \infty$ for any reach $k \in \cup\mathcal{B}_j$. The diffuse log likelihood will adjust each of these terms to prevent this divergence. We assume that $\Omega_{\infty,k}$ in a non-zero for every reach in $\cup\mathcal{B}_j$. This avoids the cases were Δ_k is independent of κ for some $k \in \cup\mathcal{B}_j$.

Adapting Durbin and Koopman (2001, p.139), we define the diffuse likelihood to be

$$\log L_d(\mathbf{y}|\theta) = \lim_{\kappa\to\infty}\left[\log L(\mathbf{y}) + \frac{n_\infty}{2}\log\kappa\right]$$

where n_∞ is defined to be the number of observations, or sum of the ranks of $\Delta_{\infty,k}$, for $k \in \cup\mathcal{B}_j$ with infinite innovation variance as $\kappa \to \infty$. As with the Exact Kalman recursions, we assume $\Delta_{\infty,k}$ is positive definite or the zero matrix.

Following Durbin and Koopman (2001, Section 7.2.2), we know from §2.5.2 that

$$\Delta_k^{-1} = \kappa^{-1}\Delta_{\infty,k}^{-1} + O(\kappa^{-2})$$

when $\Delta_{\infty,k}$ is positive definite. Furthermore, if $\Delta_{\infty,k}$ has rank p, Durbin and Koopman (2001, p.139) show that

$$\begin{aligned} -\log|\Delta_k| &= \log|\Delta_k^{-1}| = \log|\kappa^{-1}\Delta_{\infty,k}^{-1} + O(\kappa^{-2})| \\ &= -p\log(\kappa) + \log|\Delta_{\infty,k}^{-1} + O(\kappa^{-1})| \end{aligned}$$

from which we see that

$$\lim_{\kappa\to\infty}(-\log|\Delta_k| + p\log(\kappa)) = \log|\Delta_{\infty,k}^{-1}| = -\log|\Delta_{\infty,k}|$$

for $k \in \cup\mathcal{B}_j$. For these same k when $\Delta_{\infty,k}$ is positive definite, we also see that

$$\begin{aligned} \lim_{\kappa\to\infty}\mathbf{v}(k)^T\Delta_k^{-1}\mathbf{v}(k) &= \lim_{\kappa\to\infty}\left[\mathbf{v}^{(0)}(k) + \mathbf{v}^{(1)}(k) + O(\kappa^{-2})\right]^T\left[\kappa^{-1}\Delta_{\infty,k}^{-1} + O(\kappa^{-2}\right] \\ &\quad \times\left[\mathbf{v}^{(0)}(k) + \mathbf{v}^{(1)}(k) + O(\kappa^{-2})\right] \\ &= 0. \end{aligned}$$

For the alternative case that $\Delta_{\infty,k} = 0$ with $\Delta_k^{-1} = \Delta_{\star,k}^{-1} + O(\kappa^{-1})$, we have

$$\lim_{\kappa\to\infty} -\log|\Delta_k| = -\log|\Delta_{\star,k}^{-1}|$$

and

$$\begin{aligned} \lim_{\kappa\to\infty}\mathbf{v}(k)^T\Delta_k^{-1}\mathbf{v}(k) &= \lim_{\kappa\to\infty}\left[\mathbf{v}^{(0)}(k) + \kappa^{-1}\mathbf{v}^{(0)}(k) + O(\kappa^{-2}\right]^T \times\left[\Delta_{\star,k}^{-1} + O(\kappa^{-1})\right] \\ &\quad \times\left[\mathbf{v}^{(0)}(k) + \kappa^{-1}\mathbf{v}^{(0)}(k) + O(\kappa^{-2})\right] \\ &= \mathbf{v}^{(0)T}(k)\Delta_{\star,k}^{-1}\mathbf{v}^{(0)}(k). \end{aligned}$$

This leads to the diffuse log likelihood

$$\log L_d(\mathbf{y}|\theta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{k\in\cup\mathcal{B}_j} w_k - \frac{1}{2}\sum_{k\notin\cup\mathcal{B}_j}\left(\log|\Delta_k| + \mathbf{v}(k)^T\Delta_k^{-1}\mathbf{v}(k)\right) \quad (3.2)$$

where

$$w_k = \begin{cases} \log|\Delta_{\infty,k}| & \text{if } \Delta_{\infty,k} \text{ positive definite} \\ \log|\Delta_{\star,k}| + \mathbf{v}^{(0)T}(k)\Delta_{\star,k}^{-1}\mathbf{v}^{(0)}(k) & \text{if } \Delta_{\infty,k} = 0 \end{cases}$$

for any $k \in \cup\mathcal{B}_j$. Durbin and Koopman (2001, p.142) indicate that for models of interest, parameter estimates for θ obtained by maximizing $\log L(\mathbf{y}|\theta)$ for fixed κ converge to estimates of θ obtained by maximizing $\log L_d(\mathbf{y}|\theta)$. However, they do not provide insight to particular models of interest or any formal proof. They do provide an argument for a random walk plus noise model (Durbin and Koopman, 2001, p.31).

3.2 Concentrated Likelihood

Since there will exist many first order reaches in a stream network, it may be desirable to attempt to estimate these initial conditions instead of using a diffuse prior. Furthermore, a diffuse approach may seem unnatural since all observed values over a network will be finite. Depending on model assumptions, the existence of multiple first order reaches can be viewed as replications from which information about the initial conditions can be attained, in particular the mean and variance of the initial state distribution. Assume that each first order reach is a random draw from

$$\mathbf{X}(k) \sim N(\mu_0, C_0)$$

where μ_0^T is $p \times 1$ and C is $p \times p$ diagonal. If we define n_1 to be the number of first order reaches in the network, then define $\mathbf{x}_0$ to be the $(p \times n_1) \times 1$ vector of stacked first order states. Further define $\mu = J_{n_1} \otimes \mu_0$ and $C = I_{n_1} \otimes C_0$ with corresponding dimensions $(p \times n_1) \times 1$ and $(p \times n_1) \times (p \times n_1)$. The matrices F, G, and L have the form previously described in §2.4.5. Under a Gaussian model, the joint distribution of initial states and $\mathbf{Y}$ is defined by

$$\begin{bmatrix} \mathbf{x}_0 \\ \mathbf{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ GF\mu \end{bmatrix}, \begin{bmatrix} C & CF^TG^T \\ GFC & \Sigma + GFCF^TG^T \end{bmatrix} \right)$$

following the matrix form of a state-space model previously discussed. The matrix Σ is the variance of $\mathbf{Y}|\mathbf{x}_0$, and is a function of the covariance matrices of $\mathbf{W}$ and $S\mathbf{V}$ defined in §2.4.5. Recall from §2.4.5 that

$$\begin{aligned} \mathbf{v} &= (I - GK)\mathbf{y} - GL\mathbf{X}_0 \\ &= C^*\mathbf{y} - GL\mathbf{X}_0 \end{aligned}$$

where $E[\mathbf{V}] = \mathbf{0}$ with block diagonal variance $\mathrm{Var}\mathbf{v} = D$ because of orthonality in the innovations. Then for a fixed $\mathbf{X}_0$, we obtain a Cholesky decomposition $\Sigma^{-1} = C^{*T}D^{-1}C^*$. Together with $C^*GF = GL$, we use the relation

$$(A^{-1} + CB^{-1}D)^{-1} = A - AC(B + DAC)^{-1}DA \tag{3.3}$$

to express conditional means and variances in a form useful in evaluation of the likelihood considered below. First notice that

$$\begin{aligned} (\Sigma + GFCF^TG^T)^{-1} &= \Sigma^{-1} - \Sigma^{-1}GF\left(C^{-1} + F^TG^T\Sigma^{-1}GF\right)^{-1}F^TG^T\Sigma^{-1} \\ &= C^{*T}D^{-1}C^* - C^{*T}D^{-1}X\left(C^{-1} + X^TD^{-1}X\right)^{-1}X^TD^{-1}C^* \end{aligned}$$

where $X = GL = C^*GF$. We can then see the conditional expectation

$$\begin{aligned} \mathrm{E}[\mathbf{x}_0|\mathbf{Y}] &= \mu + CF^TG^T\left(\Sigma + GFCF^TG^T\right)^{-1}(\mathbf{Y} - GF\mu) \\ &= CC^{-1}\mu - CX^TD^{-1}X\mu + CX^TD^{-1}X(C^{-1} + X^TD^{-1}X)^{-1}X^TD^{-1}X\mu \\ &\quad + \left(C - CX^TD^{-1}X(C^{-1} + X^TD^{-1}X)^{-1}\right)X^TD^{-1}\mathbf{v}_0 \\ &= \left(C - CX^TD^{-1}X\left(C - (C^{-1} + X^TD^{-1}X)^{-1}X^TD^{-1}XC\right)\right)C^{-1}\mu \\ &\quad + \left(C^{-1} + X^TD^{-1}X\right)^{-1}X^TD^{-1}\mathbf{v}_0 \\ &= \left(C^{-1} + X^TD^{-1}X\right)^{-1}C^{-1}\mu + \left(C^{-1} + X^TD^{-1}X\right)^{-1}X^TD^{-1}\mathbf{v}_0 \\ &= (C^{-1} + X^TD^{-1}X)^{-1}(C^{-1}\mu + X^TD^{-1}\mathbf{v}_0) \end{aligned}$$

where $\mathbf{v}_0 = C^*\mathbf{Y}$, the component of the innovation vector that is independent of $\mathbf{x}_0$. The conditional variance is easily reformulated to be

$$\begin{aligned}
\mathrm{Var}[\mathbf{x}_0|\mathbf{Y}] &= C - CF^TG^T(\Sigma + GFCF^TG^T)^{-1}GFC \\
&= C - CX^TD^{-1}X\left(C - \left(C^{-1} + X^TD^{-1}X\right)^{-1}X^TD^{-1}XC\right) \\
&= C - CX^TD^{-1}X\left(C^{-1} + X^TD^{-1}X\right)^{-1} \\
&= (C^{-1} + X^TD^{-1}X)^{-1}
\end{aligned}$$

which is used with the conditional mean to derive a log-likelihood that is independent of $\mathbf{x}_0$.

Following the work of De Jong (1988), the log of the likelihood $p(\mathbf{Y}|\theta)$ can be expressed as

$$\log p(\mathbf{y}|\theta) = \log p(\mathbf{x}_0|\theta) + \log p(\mathbf{Y}|\mathbf{x}_0, \theta) - \log p(\mathbf{x}_0|\mathbf{y}, \theta)$$

from which it is seen that -2log of the likelihood is

$$l(\mathbf{y}|\theta) = l(\mathbf{x}_0|\theta) + l(\mathbf{y}|\mathbf{x}_0, \theta) - l(\mathbf{x}_0|\mathbf{y}, \theta)$$

where each component on the right hand side is decomposed into terms that may or may not be independent of C or $\mathbf{x}_0$.

Given $\mathbf{x}_0$, $\log p(\mathbf{Y}|\mathbf{x}_0)$ is easily obtained by executing the Kalman Filter with each first order $\mathbf{X}^p(k)$ initialized with $\mathbf{x}(k)$ with zero prediction error ($\Omega_k^p = 0$). Although the innovations $\mathbf{v}(k)$ are functions of upstream first order $\mathbf{x}(k')$, the innovation variances Δ_k are not. Furthermore, all Δ_k are independent of C since they were generated with zero variance associated with first order $\mathbf{x}(k)$. Since all random vectors are assumed normal, (3.4) reduces to

$$l(\mathbf{y}|\theta) = \log|C| + (\mathbf{x}_0 - \mu)^TC^{-1}(\mathbf{x}_0 - \mu) + \log|D| + \mathbf{v}^TD^{-1}\mathbf{v} - l(\mathbf{x}_0|\mathbf{y}\theta)$$

which is the starting point to obtaining a likelihood independent of $\mathbf{x}_0$.

The innovations $\mathbf{v}(k)$ are linear in $\mathbf{x}_0$ since $\mathbf{v} = C^*\mathbf{Y} - X\mathbf{x}_0$, hence there exists a matrix X_k such $\mathbf{v}(k) \equiv \mathbf{v}_0(k) - X_k\mathbf{x}_0$, where $\mathbf{v}_0(k)$ and X_k are the rows of $C^*\mathbf{Y}$

and $X\mathbf{x}_0$ corresponding to reach k. The vector $\mathbf{v}_0$ is also defined as the innovation constructed using an initial state of $\mathbf{0}$. Then $l(\mathbf{y}|\theta)$ can be re-written as

$$\begin{aligned}\log|C| + \log|D| + (\mathbf{x}_0-\mu)^T C^{-1}(\mathbf{x}_0-\mu) + (\mathbf{v}_0 - X\mathbf{x}_0)^T D^{-1}(\mathbf{v}_0 - X\mathbf{x}_0) - l(\mathbf{x}_0|\mathbf{y}\theta) \\ = \log|C| + \log|D| + \begin{bmatrix} \mu - \mathbf{x}_0 \\ \mathbf{v}_0 - X\mathbf{x}_0 \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} \begin{bmatrix} \mu - \mathbf{x}_0 \\ \mathbf{v}_0 - X\mathbf{x}_0 \end{bmatrix} - l(\mathbf{x}_0|\mathbf{y},\theta).\end{aligned}$$

If $\hat{\mathbf{x}}_0$ is a weighted least squares estimate after regressing $(\mu^T, \mathbf{v}_0^T)^T$ on $(I, X^T)^T$ with weighting matrix $\mathrm{diag}(C, D)$, then

$$\begin{bmatrix} \mu \\ \mathbf{v}_0 \end{bmatrix} - \begin{bmatrix} I \\ X \end{bmatrix} \mathbf{x}_0 = M \begin{bmatrix} \mu \\ \mathbf{v}_0 \end{bmatrix} - \begin{bmatrix} I \\ X \end{bmatrix} (\mathbf{x}_0 - \hat{\mathbf{x}}_0),$$

and the error projection matrix matrix $M = I - H$ for hat matrix H is

$$M = I - \begin{bmatrix} I \\ X \end{bmatrix} \left[\begin{bmatrix} I \\ X \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} \begin{bmatrix} I \\ X \end{bmatrix} \right]^{-1} \begin{bmatrix} I \\ X \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1}.$$

Using properties of M, $l(\mathbf{y}|\theta)$ can be expressed in the form of

$$\begin{aligned}\log|C| + \log|D| \quad &+ \quad \begin{bmatrix} \mu \\ \mathbf{v}_0 \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} M \begin{bmatrix} \mu \\ \mathbf{v}_0 \end{bmatrix} \\ &+ \quad (\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T \left[C^{-1} + X^T D^{-1} X\right] (\mathbf{x}_0 - \hat{\mathbf{x}}_0) - l(\mathbf{x}_0|\mathbf{y},\theta)\end{aligned}$$

where one of the exponent terms is seen to be that associated with the conditional distribution $\mathbf{x}_0|\mathbf{Y}$. By adding the appropriate term involving the determinant of the conditional variance, $l(\mathbf{y}|\theta)$ reduces to

$$\begin{aligned} l(\mathbf{y}|\theta) &= \log|C| + \log|D| - \log|(C^{-1} + X^T D^{-1} X)^{-1}| \\ &\quad + \begin{bmatrix} \mu \\ \mathbf{v}_0 \end{bmatrix}^T \begin{bmatrix} C & 0 \\ 0 & D \end{bmatrix}^{-1} M \begin{bmatrix} \mu \\ \mathbf{v}_0 \end{bmatrix} \\ &= \log|D| + \log|I + CS| + \mu^T C^{-1}\mu + \mathbf{v}_0^T D^{-1}\mathbf{v}_0 \\ &\quad -(C^{-1}\mu + \mathbf{s})^T (C^{-1} + S)^{-1}(C^{-1}\mu + \mathbf{s}) \\ &= \log|D| + \log|I + CS| + \mu_0^T A^T C^{-1} A\mu_0 + \mathbf{v}_0^T D^{-1}\mathbf{v}_0 \\ &\quad -(C^{-1}A\mu_0 + \mathbf{s})^T (C^{-1} + S)^{-1}(C^{-1}A\mu_0 + \mathbf{s}) \end{aligned}$$

where $S = X^T D^{-1} X$, $\mathbf{s} = X^T D^{-1}\mathbf{v}_0$ and $A = J_{n_1} \otimes I_m$ if m is the dimension of each first order $\mathbf{X}(k)$.

Under this formulation, the likelihood can be concentrated with respect to μ_0. Differentiating with respect to μ_0, the maximum likelihood estimator for μ_0 is

$$\begin{aligned}\hat{\mu}_0 &= (A^TC^{-1}A - A^TC^{-1}(C^{-1}+S)^{-1}C^{-1}A)^{-1}A^TC^{-1}(C^{-1}+S)^{-1}\mathbf{s} \\ &= (A^TC^{-1}(C^{-1}+S)^{-1}SA)^{-1}A^TC^{-1}(C^{-1}+S)^{-1}\mathbf{s}\end{aligned}$$

which can be substituted into $l(\mathbf{y}|\theta)$ to get

$$\begin{aligned}l(\mathbf{y}|\theta) &= \log|D| + \log|I+CS| + \mathbf{v}_0^TD^{-1}\mathbf{v}_0 \\ &\quad +\mu_0^T\left(A^TC^{-1}A - A^TC^{-1}(C^{-1}+S)^{-1}C^{-1}A\right)\mu_0 \\ &\quad -2\mu_0^TA^TC^{-1}(C^{-1}+S)^{-1}\mathbf{s} - \mathbf{s}^T(C^{-1}+S)\mathbf{s} \\ &= \log|D| + \log|I+CS| + \mathbf{v}_0^TD^{-1}\mathbf{v}_0 - \\ &\quad \mathbf{s}^T(C^{-1}+S)^{-1}C^{-1}A(A^TC^{-1}(C^{-1}+S)^{-1}SA)^{-1}A^TC^{-1}(C^{-1}+S)^{-1}\mathbf{s} \\ &\quad -\mathbf{s}^T(C^{-1}+S)\mathbf{s}.\end{aligned}$$

With this, we have a closed form expression for the maximum likelihood estimator of the mean for the initial states, whereas the variance parameters can be obtained in the numerical optimization of $l(\mathbf{y}|\theta)$.

3.3 Missing Data

In the case where data are unobserved for some reaches, a number of techniques can be used for constructing a likelihood, although some are only approximations. Substitution using expressions for a smoothed $\mathbf{Y}(k)$ when $\mathbf{Y}(k)$ is missing or an Expectation Maximization (EM) algorithm can be used to obtain approximations to the likelihood whereas an alternative state-space form can be used in order to obtain the exact likelihood. We assume that the entire vector $\mathbf{Y}(k)$ is missing, denoted by $\mathbf{Y}(k) = .$, rather than single elements. This section is dedicated to obtaining an exact form of the likelihood. An expectation maximization algorithm is presented in §3.4.

As a counterpart to Brockwell and Davis (1996, p.276), we introduce a new series $\{\mathbf{Y}^*(k)\}$, related to the original process $\{\mathbf{X}(k)\}$, by the modified observation equation

$$\mathbf{Y}^*(k) = G_k^*\mathbf{X}(k) + \mathbf{W}^*(k)$$

where

$$G_k^* = \begin{cases} G_k & \text{if } \mathbf{Y}(k) = . \\ 0 & \text{otherwise,} \end{cases} \qquad \mathbf{W}^*(k) = \begin{cases} \mathbf{W}(k) & \text{if } \mathbf{Y}(k) \neq . \\ \mathbf{N}(k) & \text{otherwise,} \end{cases}$$

where

$$\mathbf{N}(k) \sim N(\mathbf{0}, I_{n_k x n_k}),$$

$$\mathbf{N}(k) \perp \mathbf{X}(j) \text{ for all first order } j \ ,$$

$$\mathbf{N}(k) \perp \begin{bmatrix} \mathbf{V}(j) \\ \mathbf{W}(j) \end{bmatrix} \text{ for all } k, j$$

These equations constitute a state-space representation for the new series $\{\mathbf{Y}^*(k)\}$ which coincides with $\{\mathbf{Y}(k)\}$ for $\mathbf{Y}(k) \neq .$ and takes on a random value independent of the parameter vector θ when $\mathbf{Y}(k)$ is missing. Further define a new sequence $\{\mathbf{y}^*(k)\}$ by

$$\mathbf{y}^*(k) = \begin{cases} \mathbf{y}(k) & \mathbf{Y}(k) \neq . \\ 0 & \mathbf{Y}(k) = . \end{cases}$$

Under this framework, there is a direct relationship between the likelihood based on the original observed data with that based on the newly created sequence $\{\mathbf{y}^*(k)\}$, namely

$$L(\theta, \mathbf{y}) = (2\pi)^{(n_m/2)} L(\theta, \mathbf{y}^*)$$

where n_m is the number of occurrences of missing data. (Note: n_m takes into account the number of reaches as well as the dimension of the observation vector.) From this we can determine the exact likelihood based on the observed data from that based on the new sequence $\{\mathbf{y}^*(k)\}$.

3.4 Expectation-Maximization Algorithm

When the objective function is complicated by missing data, an expectation-maximization (EM) algorithm is another tool that can be used to obtain parameter estimates. Define $\mathbf{W}$ to be a "complete" data vector consisting of the *observed* $\mathbf{Y}$ and *unobserved* $\mathbf{X}$. The EM algorithm is an iterative procedure that allows computation of maximum likelihood estimates based only on the observed data $\mathbf{Y}$ (Dempster et al., 1977). Here, we follow the development of EM as given in Brockwell and Davis (1996, p.282). If $\theta^{(i)}$ is the estimate of the parameter vector after the i^{th} iteration, then the two steps in the next iteration involve calculating an expectation (E-step) with respect to the density function $f(\mathbf{x}|\mathbf{y}, \theta^{(i)})$ and maximization (M-step) of this expectation with respect to θ. The E-step is defined by the conditional expectation

$$Q(\theta|\theta^{(i)}) = E_{\theta^{(i)}}\left[l(\theta; \mathbf{X}, \mathbf{Y})|\mathbf{Y}\right] \tag{3.4}$$

where $l(\theta; \mathbf{X}, \mathbf{Y}) = \ln(f(\mathbf{x}, \mathbf{y}; \theta))$.

Paralleling that of Brockwell and Davis (1996, p.281), a brief analytical argument shows that if $\theta^{(i)}$ converges to $\hat{\theta}$, then $\hat{\theta}$ must also be the maximum likelihood estimator based on $l(\theta, \mathbf{Y})$. Using (3.4) we see that

$$Q(\theta|\theta^{(i)}) = \int (\ln f(\mathbf{x}|\mathbf{Y}; \theta)) f(\mathbf{x}|\mathbf{Y}; \theta^{(i)}) d\mathbf{x} + l(\theta; \mathbf{Y})$$

If $\theta^{(i)}$ is converging to $\hat{\theta}$, then using the fact that $Q'(\theta^{(i+1)}|\theta^{(i)}) = 0$ and letting $i \to \infty$, we see that

$$\begin{aligned}
0 &= \lim_{i\to\infty}\left\{ \int \frac{\frac{\partial}{\partial\theta} f(\mathbf{x}|\mathbf{Y}; \theta)_{\theta=\theta^{(i+1)}}}{f(\mathbf{x}|\mathbf{Y}; \theta)_{\theta=\theta^{(i+1)}}} f(\mathbf{x}|\mathbf{Y}; \theta^{(i)}) d\mathbf{x} + l'(\theta^{(i+1)}; \mathbf{Y}) \right\} \\
&= \int \frac{\frac{\partial}{\partial\theta} f(\mathbf{x}|\mathbf{Y}; \theta)_{\theta=\hat{\theta}}}{f(\mathbf{x}|\mathbf{Y}; \theta)_{\theta=\hat{\theta}}} f(\mathbf{x}|\mathbf{Y}; \hat{\theta}) d\mathbf{x} + l'((\hat{\theta}); \mathbf{Y}) \\
&= \int \frac{\partial}{\partial\theta} f(\mathbf{x}|\mathbf{Y}; \hat{\theta}) d\mathbf{x} + l'(\hat{\theta}; \mathbf{Y})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial\theta}\int f(\mathbf{x}|\mathbf{Y};\hat{\theta})d\mathbf{x} + l^{'}(\hat{\theta};\mathbf{Y}) \\
&= \frac{\partial}{\partial\theta}(1) + l^{'}(\hat{\theta};\mathbf{Y})
\end{aligned}
$$

showing that $\hat{\theta}$ must also be a solution to the likelihood equations $l^{'}(\theta,\mathbf{Y}) = 0$.

In the case where $\mathbf{X}$ is a vector of missing observations such that the exact likelihood is difficult to obtain and maximize, the EM approach may have a large computational advantage. For $\mathbf{W} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N(\mathbf{0}, \Sigma(\theta))$, the log-likelihood given by the complete data is

$$
l(\theta;\mathbf{W}) = -\frac{n}{2} - \frac{1}{2}\ln\det\Sigma(\theta) - \frac{1}{2}\mathbf{W}^T\Sigma^{-1}(\theta)\mathbf{W} \tag{3.5}
$$

where n is the number of observations in the complete data vector. Using properties of the Multivariate Normal distribution, it is straightforward to show that

$$
\mathbf{W}|\mathbf{Y} \sim \mathrm{N}\left(\begin{bmatrix} \hat{\mathbf{X}} \\ \mathbf{Y} \end{bmatrix}, \begin{bmatrix} \Sigma_{X|Y}(\theta) & 0 \\ 0 & 0 \end{bmatrix}\right)
$$

which is needed in the E-step when taking the expectation of (3.5) with respect to $f(\mathbf{x}|\mathbf{Y};\theta^{(i)})$. The calculation involves taking the expectation of the quadratic form $\mathbf{W}^T\Sigma^{-1}(\theta)\mathbf{W}$, which is found to be

$$
\begin{aligned}
E_{\theta^{(i)}}\left[\mathbf{W}^T\Sigma(\theta)\mathbf{W}|\mathbf{Y}\right] &= \mathrm{Trace}\left(\Sigma^{-1}(\theta)\Sigma(\theta^{(i)})\right) + E_{\theta^{(i)}}\left[\mathbf{W}|\mathbf{Y}\right]^T\Sigma^{-1}(\theta)E_{\theta^{(i)}}\left[\mathbf{W}|\mathbf{Y}\right] \\
&= \mathrm{Trace}\left(\begin{bmatrix} \Sigma_{xx}(\theta) & \Sigma_{xy}(\theta) \\ \Sigma_{yx}(\theta) & \Sigma_{yy}(\theta) \end{bmatrix}^{-1}\begin{bmatrix} \Sigma_{x|y}(\theta^{(i)}) & 0 \\ 0 & 0 \end{bmatrix}\right) \\
&\quad + \hat{\mathbf{W}}^T\Sigma^{-1}(\theta)\hat{\mathbf{W}} \\
&= \mathrm{Trace}\left(\Sigma_{x|y}^{-1}(\theta)\Sigma_{x|y}(\theta^{(i)})\right) + \hat{\mathbf{W}}^T\Sigma^{-1}(\theta)\hat{\mathbf{W}}
\end{aligned}
$$

since the first block entry of $\Sigma^{-1}(\theta)$ is $\Sigma_{x|y}^{-1}(\theta) = (\Sigma_{xx}(\theta) - \Sigma_{xy}(\theta)\Sigma_{yy}^{-1}(\theta)\Sigma_{yx}(\theta))^{-1}$. With this, it is clear that

$$
Q(\theta|\theta^{(i)}) = l(\theta;\hat{\mathbf{W}}) - \frac{1}{2}\mathrm{Trace}\left(\Sigma_{x|y}^{-1}(\theta)\Sigma_{x|y}(\theta^{(i)})\right). \tag{3.6}
$$

This application of the EM algorithm uses the Kalman recursions developed for the stream network. The E-step entails estimation of the missing observations using

their *smoothed* values from the downstream filter and upstream smoother in order to obtain $Q(\theta|\theta^{(i)})$, which is then maximized to obtain $\theta^{(i+1)}$. When the increments $\theta^{(i+1)} - \theta^{(i)}$ are small, the second term in (3.6) can be ignored, further simplifying the estimation (see Brockwell and Davis, 1996, p.282).

3.5 Simulated Example

Consider a relatively small stream network, or only a small region of a larger stream network as depicted in Figure 2.2. A univariate model is examined where the contributions from the upstream reaches as well as a conditional variances are defined to be functions of reach orders. This model is believed to be reasonable since reach order has been shown to be proportional to relative watershed dimensions, channel size, and stream discharge at that place in the stream system.

If i, j, and k are the orders of u_1, u_2, and k respectively, then the state-space model defined by (2.1) and (2.2) is

$$Y(k) = X(k) \tag{3.7}$$

$$X(k) = \phi_{k,i}X(u_1) + \phi_{k,j}X(u_2) + V(k) \tag{3.8}$$

where $G_k = 1$, $W(k) = 0$, and $Q_k = \sigma^2_{ij}$, with initial assumptions that $X(j) = 0$ and $V(j) \sim N(0, \tau^2)$ for any first order reach j. The recursions are defined as

$$\begin{aligned} Y(3) &= \phi_{21}Y(1) + \phi_{21}Y(2) + V(3) \\ Y(5) &= \phi_{22}Y(3) + \phi_{21}Y(4) + V(5) \\ Y(7) &= \phi_{22}Y(5) + \phi_{21}Y(6) + V(7). \end{aligned}$$

Under this model, the parameter vector $\theta = [\phi_{21}, \phi_{22}, \sigma^2_{11}, \sigma^2_{21}, \tau^2]^T$, hence two ϕ parameters and three variance parameters to estimate based on only seven observations. As may be the case in many parametric models on stream networks, there are a relatively small number of observations that are actually used in estimation of each parameter.

In order to construct a likelihood in terms of the innovations and prediction error variances, the Kalman Filter is applied sequentially over each segment. To initialize the filter, the predicted values for all first order streams are defined to be zero, with prediction error variance τ^2. When there are no missing observations, the log of the likelihood defined by (3.1) is easily obtained through execution of the Kalman Filter over this small network. The filter is applied sequentially over the sets $\mathcal{S}_1 = [1, 2, 4, 6]$ followed by $\mathcal{S}_2 = [\{3, 5, 7\}]$.

The innovations are defined by $I(k) = Y(k)$ with $\Delta_k = \tau^2$ for any $k \in \mathcal{S}_1$. Starting with reach 3, we see that

$$\begin{aligned} I(3) &= Y(3) - \phi_{21}Y(1) - \phi_{21}Y(2) \\ I(5) &= Y(5) - \phi_{22}Y(3) - \phi_{21}Y(4) \\ I(7) &= Y(7) - \phi_{22}Y(5) - \phi_{21}Y(6) \end{aligned}$$

where $\Delta_3 = \sigma_{11}^2$ and $\Delta_5 = \Delta_7 = \sigma_{21}^2$. With that, the Log-likelihood is easily constructed as

$$\begin{aligned} l(\theta, \mathbf{Y}) \propto & -\ln(\sigma_{21}^2) - \frac{1}{2}\ln(\sigma_{11}^2) - 2\ln(\tau^2) \\ & -\frac{1}{2}\left(\frac{I(7)^2 + I(5)^2}{\sigma_{21}^2} + \frac{I(3)^2}{\sigma_{11}^2} + \frac{I(1)^2 + I(2)^2 + I(4)^2 + I(6)^2}{\tau^2}\right) \end{aligned}$$

To obtain parameter estimates, we use the Log-likelihood as a criterion function to optimize with respect to the unknown quantities of interest. When missing data are present, an exact likelihood can still be obtained via the alternative state-space form described in §3.3. Suppose that the observation on reach 5 was unknown. The observation equation can be modified by

$$G_k^* = \begin{cases} G_k & \text{if } k \neq 5 \\ 0 & \text{otherwise,} \end{cases} \qquad W^*(k) = \begin{cases} W(k) & \text{if } t \neq 5 \\ N(k) & \text{otherwise,} \end{cases}$$

where

$$N(5) \sim N(0,1), \quad N(5) \perp X(k) \text{ for k} \in \mathcal{S}_1, \quad N(5) \perp \begin{bmatrix} V(k) \\ W(k) \end{bmatrix}.$$

The new series $\{Y^*(k)\}$ coincides with $\{Y(k)\}$ for $k \neq 5$ and takes on a random value independent of the parameter vector θ when $k = 5$. Define a new sequence $\{y^*(k)\}$ by

$$y^*(k) = \begin{cases} y(k) & k \neq 5 \\ 0 & k = 5 \end{cases}$$

where $L(\theta, \mathbf{y}) = (2\pi)^{(1/2)} L(\theta, \mathbf{y}^*)$. The one-step predictors of $\{Y^*(t)\}$ and their error covariance matrices Δ_t^* are now

$$\begin{aligned} I^*(3) &= Y^*(3) - \phi_{21} Y^*(1) - \phi_{21} Y^*(2) \\ I^*(5) &= 0 \\ I^*(7) &= Y^*(7) - \phi_{22}(\phi_{22} Y^*(3) + \phi_{21} Y^*(4)) - \phi_{21} Y^*(6) \end{aligned}$$

where $\Delta_3^* = \sigma_{11}^2$, $\Delta_5^* = 1$, and $\Delta_7^* = \phi_{22}^2 \sigma_{21}^2 + \sigma_{21}^2$. The resulting Log-likelihood is

$$\begin{aligned} l(\theta, \mathbf{Y}) \quad \propto \quad & -\frac{1}{2}\ln(\sigma_{21}^2) - \frac{1}{2}\ln(\phi_{22}^2 + 1) - \frac{1}{2}\ln(\sigma_{11}^2) - 2\ln(\tau^2) \\ & -\frac{1}{2}\left(\frac{I^*(7)^2}{\sigma_{21}^2(\phi_{22}^2 + 1)} + \frac{I^*(3)^2}{\sigma_{11}^2} + \frac{I^*(1)^2 + I^*(2)^2 + I^*(4)^2 + I^*(6)^2}{\tau^2} \right) \end{aligned}$$

which we see is slightly more complicated than the function with no missing data. However, this may not be the case with larger networks, a more complicated model, and more missing data.

Alternatively, the EM algorithm from §3.4 can be used to find an approximation to the likelihood where parameter estimates and smoothed predictions are obtained in an iterative fashion. Let $\mathbf{Y}$ denote the vector of observed data. We can define the steps of the EM algorithm by

E-step

1. Derive smoothed estimate $Y^s(5)$ for the missing observation given $\theta^{(i)}$
2. Calculate $Q(\theta|\theta^{(i)}) = l(\theta, Y^s(k), \mathbf{Y}) - \left(\frac{1}{2}\right)\frac{\phi_{22}^2+1}{\phi_{22}^{(i)2}+1}$
3. Calculate $\widetilde{Q}(\theta|\theta^{(i)}) = l(\theta, Y^s(k), \mathbf{Y})$

M-step

1. Numerically maximize $Q(\theta|\theta^{(i)})$ and $\widetilde{Q}(\theta|\theta^{(i)})$ with respect to θ to obtain $\theta^{(i+1)}$.

This process is iterated until $\| \theta^{(i+1)} - \theta^{(i)} \|^2$ is within a specified tolerance.

A simulation was performed to evaluate the likelihood methods of estimation dicussed in this chapter. Data were generated for the above network with the true parameter vector $\theta = [.4, .6, 9, 4, 8]^T$. Initially drawing data for the first order streams, the data for reaches 3, 5, and 7 were computed sequentially. Adding more reaches only complicates this model with more parameters, so observations for independent trees were generated in order to increase the number of observations for estimation. The effect of missing data is determined by setting the value of reach 5 in Figure 2.2 to be missing. Configurations of 1, 10, and 100 trees were considered to see the extent that sample size influences parameter estimation. Furthermore, various degrees of missing values are considered setting $Y(5)$ to missing for selected trees.

Maximum Likelihood Estimation (MLE) is performed to obtain information about the parameter vector given the available data. For simplicity, the variance parameters $\sigma_{11}^2, \sigma_{21}^2$, and τ^2 are assumed to be known, so the parameter vector to be estimated is $\theta = [\phi_{21}, \phi_{22}]^T$. Initial values for numerical optimization were arbitrarily chosen to be [0.3,0.5]. Other points in a neighborhood of [0.3,.05] were considered initially, but preliminary results were not dependent on these starting values.

Parameter estimates were obtained using the "full" likelihood if no data were missing, whereas the exact likelihood was maximized when missing data were

Table 3.1: Maximum Likelihood Estimates for $\theta = [\phi_{21}, \phi_{22}]^T = [0.4, 0.6]^T$

Trees		Proportion Observed[1]	$\hat{\phi_{21}}$[2]	rmse($\hat{\phi_{21}}$)	bias($\hat{\phi_{21}}$)	$\hat{\phi_{22}}$	rmse($\hat{\phi_{22}}$)	bias($\hat{\phi_{22}}$)
1	Full	1	0.5304	0.923	0.130	0.6152	1.162	0.015
	Exact	0.5	0.6368	1.303	0.237	0.3643	0.562	-0.236
	EM	0.5	0.6183	1.149	0.218	0.4017	0.546	-0.198
	AEM	0.5	0.6427	1.492	0.243	0.9564	1.841	0.356
2	Full	1	0.4038	0.355	0.004	0.5410	0.439	-0.059
	Exact	0.5	0.4857	0.516	0.086	0.3746	0.509	-0.225
	EM	0.5	0.4857	0.516	0.086	0.3757	0.507	-0.224
	AEM	0.5	0.4687	0.532	0.069	0.6638	0.575	0.064
10	Full	1	0.419	0.143	0.019	0.5695	0.135	-0.031
	Exact	0.5	0.4303	0.180	0.030	0.5230	0.230	-0.077
	EM	0.5	0.4303	0.180	0.030	0.5231	0.230	-0.077
	AEM	0.5	0.4164	0.168	0.016	0.7008	0.183	0.101
100	Full	1	0.4076	0.036	0.008	0.5941	0.041	-0.006
	Exact	0.5	0.4124	0.047	0.012	0.5911	0.048	-0.009
	EM	0.5	0.4124	0.047	0.012	0.5911	0.048	-0.009
	AEM	0.5	0.3989	0.047	-0.001	0.7111	0.118	0.111

[1] Fraction of observed values on reaches formed by a second order reach merging with a first order reach.

[2] Estimates, rmse, and bias empirically determined from 95 of 100 simulated realizations.

present. Parameter estimates for the different number of trees per network are shown in Table 3.1. Numerical issues resulted in unusual parameter estimates in a small number of simulations, hence empirical estimates for parameters, bias and root mean square error are determined from a trimmed set of 95 out of 100 simulations of each network.

As an alternative to maximizing the exact likelihood, the EM algorithm previously introduced was used under two formulations for $Q(\theta|\theta^{(i)})$. Initially, (3.6) was used for estimation, followed by an "adjusted" formulation where the second term in (3.6) was ignored. In both cases, *complete* data were obtained by using the smoothed estimate for the missing values via the Kalman recursions. Starting with initial estimate for θ, the original data were smoothed, and new estimates were obtained by maximizing (3.6) given this newly created data vector. Subsequent estimates were found by re-smoothing the original data using $\hat{\theta}$ from the previous iteration. The algorithm was iterated until parameter estimates stabilized. Additional estimates were obtained in the same fashion using the "adjusted" form of (3.6). Results from these approaches are shown in Table 3.1.

This relatively simple network possesses computational challenges, which were overcome by creating replicates. With a small number of observations, the numerical optimization of the likelihood can be problematic, as was the case in this example with only one replication. Here, the innovation for reach 3 seemed to dominate in the maximization, which resulted in $\hat{\phi}_{21} \approx \frac{y(3)}{y(1)+y(2)}$. Because of this, the information about ϕ_{21} from reaches 4 and 6 was essentially ignored. Given the nature and ordering of stream networks, numerical dilemmas are likely to occur as more parameters are modeled relative to the amount of observed data.

The exact likelihood was easy to obtain algebraically resulting in optimization without computational difficulties. Although the EM algorithms possessed no advantage, they provided illustration for its potential use. Using the Kalman Filter

to construct the exact likelihood is also an option, however, this may add to the computational burden since the Kalman Filter must be run with each iteration in the numerical optimization. Even in the cases with small numbers of trees, this option resulted in unreasonable computational requirements.

Although point estimates are easily obtained, properties such as bias and mean square error remain as open areas of research.

Chapter 4

AUTOREGRESSIVE MOVING AVERAGES

The autoregressive moving average models, or ARMA models, specify a large class of models defined by stochastic linear difference equations with constant coefficients. Properties of this parametric family of stationary processes in time series are established in Brockwell and Davis (1991, Chapter 3). In this chapter, we extend these results from time series to processes over the tree structure of a stream network.

4.1 ARMA Processes

Definition 4.1.1 The process $\{X(k)\}$ on a stream network is an autoregressive moving average, or ARMA(p, q), process if it is (weakly) stationary and if for some real constants $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$,

$$X(k) - \phi_1 BX(k) - \cdots - \phi_p B^p X(k) = Z(k) + \theta_1 BZ(k) + \cdots + \theta_q B^q Z(k) \quad (4.1)$$

for every reach k, where $Z(k) \sim WN(0, \sigma^2)$.

In a more compact form, this can be written as

$$\phi(B)X(k) = \theta(B)Z(k) \quad (4.2)$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are polynomials of degree p and q with no common factors, and are defined by

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

Parameter constraints are obtained through the concepts of *causality* and *invertibility* of an ARMA process.

4.2 Causality and Invertibility

The concepts of *causality* and *invertibility* (e.g. Brockwell and Davis, 1991) are considered, and involve defining a relationship between a network process $X(k)$ and the white noise process $Z(k)$ in terms of infinite sums. These concepts over the stream network are defined in terms of mean square convergence as opposed to the more restrictive absolute convergence. Although the natural branching that occurs in the stream network makes writing these infinite sums notationally complex, they are obtainable and simplified using lag notation. The following results are extensions of ARMA(p, q) processes in time. The results and proofs herein parallel those presented in Brockwell and Davis (1991, Chapter 3).

Definition 4.2.1 Causality

An ARMA(p, q) process defined by (4.2) on a stream network is said to be a causal function of $Z(k)$ if there exists a sequence of constants $\{\psi_i\}$ where $\sum_{i=0}^{\infty} |\psi_i| \sqrt{2}^{-i} < \infty$ such that

$$X(k) = \psi(B)Z(k) = \sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k), \tag{4.3}$$

the series converging in mean square, where $Z_{ij}(k)$ is from the j^{th} ancestral reach of the i^{th} generation upstream of k.

Proposition 4.2.1 *Suppose $X(k)$ is a zero mean stationary process over a stream network with autocovariance function $\gamma_X(\cdot)$ such that $\sum_{h=0}^{\infty} |\gamma_X(h)| \sqrt{2}^{h} < \infty$ and*

$X(k) \perp X(k')$ when k and k' are not flow connected. If $\sum_{i=0}^{\infty} \psi_i^2 2^{-i} < \infty$, then the series

$$\psi(B)X(k) = \sum_{i=0}^{\infty} \psi_i B^i X(k) = \sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} X_{ij}(k)$$

converges in mean square. Furthermore, the mean square limit $Y(k)$ is a stationary process with autocovariance function

$$\gamma_Y(0) = \sum_{i=0}^{\infty} \frac{\psi_i^2}{2^i} \gamma_X(0) + 2 \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \frac{\psi_j \psi_{j+i}}{2^i} \gamma_X(i)$$

and for $h > 0$,

$$\begin{aligned} \gamma_Y(h) = & \sum_{i=0}^{h-1} \frac{\psi_i}{2^i} \sum_{j=0}^{\infty} \psi_j \gamma_X(h-i+j) + \sum_{i=0}^{\infty} \frac{\psi_{i+h}}{2^{i+h}} \psi_i \gamma_X(0) \\ & + \sum_{h'=1}^{\infty} \sum_{i=0}^{\infty} \frac{\psi_{i+h}}{2^{i+h}} \psi_{i+h'} \gamma_X(h') + \sum_{h'=1}^{\infty} \sum_{i=0}^{\infty} \frac{\psi_{i+h+h'}}{2^{i+h}} \psi_i \gamma_X(h') \end{aligned} \tag{4.4}$$

Proof First show convergence in mean square. For $n > m > 0$,

$$\begin{aligned} E\left| \sum_{i=m}^{n} \psi_i \sum_{j=1}^{2^i} \frac{X_{ij}(k)}{2^i} \right|^2 &= \sum_{i=m}^{n} (\frac{\psi_i}{2^i})^2 2^i \gamma_X(0) + 2 \sum_{h=1}^{n-m} \sum_{i=m}^{n-h} \frac{\psi_i \psi_{i+h}}{2^i 2^{i+h}} 2^{i+h} \gamma_X(h) \\ &= \sum_{i=m}^{n} \frac{\psi_i^2}{2^i} \gamma_X(0) + 2 \sum_{h=1}^{n-m} \gamma_X(h) \sum_{i=m}^{n-h} \frac{\psi_i \psi_{i+h}}{2^i} \\ &\leq \sum_{i=m}^{n} \frac{\psi_i^2}{2^i} \gamma_X(0) + 2 \sum_{h=1}^{n-m} |\gamma_X(h)| \left(\sum_{i=m}^{n-h} \frac{\psi_i^2}{2^i} \right)^{1/2} \left(\sum_{i=m}^{n-h} \frac{\psi_{i+h}^2}{2^i} \right)^{1/2} \\ &\leq \sum_{i=m}^{n} \frac{\psi_i^2}{2^i} \gamma_X(0) + 2 \sum_{h=1}^{\infty} |\gamma_X(h)| \left(\sum_{i=m}^{\infty} \frac{\psi_i^2}{2^i} \right)^{1/2} \left(2^h \sum_{i=m}^{\infty} \frac{\psi_{i+h}^2}{2^{i+h}} \right)^{1/2} \\ &\leq \sum_{i=m}^{n} \frac{\psi_i^2}{2^i} \gamma_X(0) + 2 \sum_{h=1}^{\infty} \sqrt{2}^h |\gamma_X(h)| \sum_{i=m}^{\infty} \frac{\psi_i^2}{2^i} \\ &= \gamma_X(0) \sum_{i=m}^{n} \frac{\psi_i^2}{2^i} + 2 \left(\sum_{h=1}^{\infty} \sqrt{2}^h |\gamma_X(h)| \right) \sum_{i=m}^{\infty} \frac{\psi_i^2}{2^i} \\ &\to 0 \text{ as } m, n \to \infty, \end{aligned}$$

since $\sum_{j=0}^{\infty} \psi_j^2 2^{-j} < \infty$ and the sum over $\sqrt{2}^h \gamma_X(h)$ is bounded. Thus, by the Cauchy Criterion, $\psi(B)X(k)$ converges in mean square, and is well defined. Let $Y(k)$ be the mean square limit of $\psi(B)X(k)$.

Using continuity of the inner product (Brockwell and Davis, 1991, p.45), we see that

$$\begin{aligned} E[Y(k)] &= \lim_{n\to\infty} \sum_{i=0}^{n} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} E[X_{ij}(k)] \\ &= 0 \end{aligned}$$

which does not depend on location k. Cross-product terms involved in determining closed form expressions for second moments can be confusing. Using independence when reaches are not flow connected will simplify many expressions. To begin, first note that

$$E\left(\sum_{i=0}^{\infty} \frac{\psi_{i+h}}{2^{i+h}} \sum_{j=1}^{2^i} X_{ij}(k')\right)\left(\sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} X_{ij}(k')\right)$$

reduces to

$$\sum_{i=0}^{\infty} \frac{\psi_{i+h}\psi_i}{2^{i+h}} \gamma_X(0) + \sum_{h'=1}^{\infty}\sum_{i=0}^{\infty} \frac{\psi_{i+h}}{2^{i+h}} \psi_{i+h'} \gamma_X(h') + \sum_{h'=1}^{\infty}\sum_{i=0}^{\infty} \frac{\psi_{i+h+h'}}{2^{i+h}} \psi_i \gamma_X(h') \qquad (4.5)$$

when the expectation is carried through the double sum. It is clear from the tree structure that the coefficients for each $\gamma_X(h')$ will depend on the offset h. Additionally, the number of occurrences of $\gamma_X(h')$ will depend on h'. The coefficients for $\gamma_X(0)$ are straightforward. Those for each $\gamma_X(h'), h \geq 1$ result from first working *upstream* from k' followed by working *downstream* to k'.

The variance in $Y(k)$ follows immediately by letting $h = 0$ in (4.5) which leads to (4.4). For covariances, we can decompose the $Y(k) = \sum_{i=0}^{\infty} \psi_i 2^{-i} \sum_{j=1}^{2^i} X_{ij}(k)$ into sums over three disjoint sets of reaches: the intermediary reaches downstream of k' that connect k with k', reaches upstream of and including k' (which are also flow connected to k), and those reaches not flow connected to k'.

Let $j^\star$ identify the appropriate reach that is downstream from and flow connected to k' at each lag i, $i = 0 \ldots h-1$ with respect to k. For $k' < k$, the covariance can be simplified using flow connectivity and expressed by

$$\mathrm{Cov}[Y(k), Y(k')] = \mathrm{Cov}\left(\sum_{i=0}^{h-1} \frac{\psi^i}{2^i} X_{i,j^\star}(k) + \sum_{i=0}^{\infty} \frac{\psi_{i+h}}{2^{i+h}} \sum_{j=1}^{2^i} X_{i,j}(k'), \sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} X_{i,j}(k') \right)$$

where the first sum corresponds to the intermediary reaches (including k), and the second sum corresponds to the reaches common to both infinite sums. All other reaches are not flow connected to both k and k', therefore independent of any reach upstream of k'.

The covariance associated with intermediary reaches is defined by

$$\begin{aligned} \mathrm{Cov}\left(\sum_{i=0}^{h-1} \frac{\psi^i}{2^i} X_{i,j^\star}(k), \sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} X_{i,j}(k') \right) &= \sum_{i=0}^{h-1} \frac{\psi^i}{2^i} \sum_{j=0}^{\infty} \frac{\psi_j}{2^j} 2^j \gamma_X(h-i+j) \\ &= \sum_{i=0}^{h-1} \frac{\psi_i}{2^i} \sum_{j=0}^{\infty} \psi_j \gamma_X(h-i+j) \quad (4.6) \end{aligned}$$

since there are 2^j reaches flow connected to $X_{i,j^\star}$ for each upstream lag j with respect to $X(k')$.

The covariances associated with reaches common to both infinite sums are not as obvious, but follow directly from (4.5). Letting $h = |k - k'|$, the covariance from these reaches is defined by (4.5), which when combined with (4.6) results in (4.4). Since the first and second moments of $Y(k)$ are independent of reach k, $Y(k)$ is weakly stationary. ∎

Remark 4.2.1 The conditions $\sum_{h=0}^{\infty} |\gamma(h)| \sqrt{2}^h < \infty$ and $\sum_{i=0}^{\infty} \psi_i^2 2^{-i} < \infty$ are conditions used to guarantee mean square convergence. Other conditions exist, such as $\sum_{i=0}^{\infty} |\psi_i| < \infty$, allowing a less stringent condition on the summability of $\gamma_X(\cdot)$. The condition we adopted for $\{\psi_j\}$ is very weak, allowing for a large class of ARMA models.

The following theorem provides the necessary and sufficient conditions for which $X(k)$ is causal, allowing expression of $X(k)$ in terms of (infinitely many) upstream $Z(k)$. The results and its proof parallel Theorem 3.1.1 in Brockwell and Davis (1991, p.85).

Theorem 4.2.1 *Let $\{X(k)\}$ be an ARMA(p,q) process over a stream network such that $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeros. Then $\{X(k)\}$ is causal if and only if $\phi(z/\sqrt{2}) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\{\psi_j\}$ in (4.3) are determined by the relation*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z), \quad |z| \leq \frac{1}{\sqrt{2}} \tag{4.7}$$

Proof For sufficient conditions for causal $X(k)$, assume $\phi(z/\sqrt{2}) \neq 0$ if $|z| \leq 1$. We set out to obtain a set of coefficients ξ_i that meet the conditions of Proposition 4.2.1.

By assumption, there exists $\epsilon > 0$ such that $1/\phi(z/\sqrt{2})$ has power series expansion

$$\sum_{i=0}^{\infty} \xi_i \left(\frac{z}{\sqrt{2}}\right)^i = \sum_{i=0}^{\infty} \frac{\xi_i}{\sqrt{2}^i} z^i$$

for $|z| < 1 + \epsilon$ (Brockwell and Davis, 1991, p.85). Thus, we see that $\xi_i \sqrt{2}^{-i}(1 + \epsilon/2)^i \to 0$ as $i \to \infty$, which implies there exists $K \in [0, \infty)$ such that

$$\frac{|\xi_i|}{\sqrt{2}^i} < \frac{K}{(1+\epsilon/2)^i} \quad \text{for } i = 0, 1, \ldots.$$

Hence, we see that $\sum_{i=0}^{\infty} |\xi_i| \sqrt{2}^{-i}$ is finite and find ξ_i through the relation

$$\xi(z) = \sum_{i=0}^{\infty} \xi_i z^i = \frac{1}{\phi(z)}, \quad |z| \leq \frac{1}{\sqrt{2}}.$$

which exists since

$$\sum_{i=0}^{\infty} \xi_i z^i \leq \sum_{i=0}^{\infty} |\xi_i z^i|$$

$$\begin{aligned} &\leq \sum_{i=0}^{\infty} |\xi_i| \left(\frac{1}{\sqrt{2}}\right)^i \\ &< \infty \end{aligned}$$

by assumption. From this we see that $\xi(z)\phi(z) \equiv 1$ for $|z| \leq \sqrt{2}^{-1}$, and define the operator $\xi(B)$.

The second criterion to utilize Proposition 4.2.1 is the covariance constraint. Again following Brockwell and Davis (1991, p.85), notice that the covariance associated with $\phi(B)X(k)$ is that corresponding to $\theta(B)Z(k)$ where $Z(k)$ are white noise. Since $\theta(B)$ has finite order q, then for any $h > q$, $\gamma_{\theta(B)Z(k)}(h) = 0$ and

$$\sum_{h=0}^{\infty} \gamma_{\theta(B)Z(k)}(h)\sqrt{2}^h = \sum_{h=0}^{q} \gamma_{\theta(B)Z(k)}(h)\sqrt{2}^h$$

which is finite when $Z(k)$ has finite variance. Then by Proposition 4.2.1 we can apply the operator $\xi(B)$ to both sides of $\phi(B)X(k) = \theta(B)Z(k)$ to find that

$$\xi(B)\theta(B)Z(k) = \psi(B)Z(k) = \sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k) = X(k)$$

where the sequence $\{\psi_i\}$ is determined by (4.7).

For necessary conditions, assume that $\{X(k)\}$ is causal, indicating $X(k)$ is the mean square limit of $\psi(B)Z(k)$ with $\sum_{i=0}^{\infty} |\psi_i|\sqrt{2}^{-i} < \infty$. Then

$$\theta(B)Z(k) = \phi(B)X(k) = \phi(B)\psi(B)Z(k).$$

Defining $\eta(z)$ to be

$$\begin{aligned} \eta(z) &= \phi(z)\psi(z) \\ &= \sum_{i=0}^{\infty} \eta_i z^i, \quad |z| \leq \frac{1}{\sqrt{2}}, \end{aligned}$$

we see from $\theta(B)Z(k) = \eta(B)Z(k)$ that

$$\sum_{i=0}^{q} \frac{\theta_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k) = \sum_{i=0}^{\infty} \frac{\eta_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k).$$

Multiplying each side by $Z_{i'j'}$ and taking expectations, we see that $\theta_i = \eta_i$ for $i = 0, \ldots, q$, and $\eta_i = 0$ for $i > q$. With that, we have

$$\begin{aligned} \theta(z) &= \eta(z) \\ &= \phi(z)\psi(z), \quad |z| \leq \frac{1}{\sqrt{2}}. \end{aligned}$$

Now, if $\phi(z) = 0$ and $|\psi(z)| < \infty$ for $|z| \leq \sqrt{2}^{-1}$, then $\theta(z) = 0$, hence $\phi(\cdot)$ and $\theta(\cdot)$ would have a common zero contradicting the assumption of no common zeros. Thus, $\phi(z) \neq 0$ for any $|z| \leq \frac{1}{\sqrt{2}}$, which is equivalent to $\phi(z/\sqrt{2}) \neq 0$ for any $|z| \leq 1$.

∎

Parameter constraints on θ are defined through the concept of *invertibility* of a process on the stream network.

Definition 4.2.2 Invertibility

The process $X(k)$ on a stream network is said to be invertible if there exists a sequence of constants $\{\pi_i\}$ such that $\sum_{i=0}^{\infty} |\pi_i|/\sqrt{2^i} < \infty$ and

$$\pi(B)X(k) = \sum_{i=0}^{\infty} \frac{\pi_i}{2^i} \sum_{j=1}^{2^i} X_{ij}(k) = Z(k), \tag{4.8}$$

the series converging in means square, where $X_{ij}(k)$ is from the j^{th} ancestral reach of the i^{th} generation upstream of k.

The following theorem identifies necessary and sufficient conditions for $\{X(k)\}$ to be invertible. The results and its proof parallel Theorem 3.1.2 in Brockwell and Davis (1991, p.86).

Theorem 4.2.2 *Let $\{X(k)\}$ be an ARMA(p,q) process over a stream network such that $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeros. Then $\{X(k)\}$ is invertible if and only if the smallest root of $\theta(z/\sqrt{2})$ is outside the unit circle, ie, $\theta(z/\sqrt{2}) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. The coefficients $\{\pi_j\}$ in (4.8) are determined by the relation*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \phi(z)/\theta(z), \quad |z| \leq \frac{1}{\sqrt{2}} \tag{4.9}$$

Proof For sufficient conditions for invertible $X(k)$, assume $\theta(z/\sqrt{2}) \neq 0$ if $|z| \leq 1$. Then there exists $\epsilon > 0$ such that $1/\theta(z/\sqrt{2})$ has power series expansion

$$\sum_{i=0}^{\infty} \eta_i \left(\frac{z}{\sqrt{2}}\right)^i = \sum_{i=0}^{\infty} \frac{\eta_i}{\sqrt{2}^i} z^i, \quad |z| < 1 + \epsilon.$$

Furthermore, since $\sum_{i=0}^{\infty} \frac{|\eta_i|}{\sqrt{2}^i} < \infty$ and the covariance associated with $\theta(B)Z(k)$ is zero for any $h > q$, we use Proposition 4.2.1 to apply $\eta(B)$ to both sides of $\phi(B)X(k) = \theta(B)Z(k)$ so that

$$\eta(B)\theta(B)Z(k) = \eta(B)\phi(B)X(k) = \pi(B)X(k) = \sum_{i=0}^{\infty} \frac{\pi_i}{2^i} \sum_{j=1}^{2^i} X_{ij}(k) = Z(k)$$

The sequence $\{\pi_i\}$ is determined by (4.9).

For necessary conditions, assume that $\{X(k)\}$ is invertible, indicating $Z(k)$ is the mean square limit of $\pi(B)X(k)$ for some sequence $\{\pi_i\}$with $\sum_{i=0}^{\infty} \frac{|\pi_i|}{\sqrt{2}^i} < \infty$. Since $\phi(B)$ is of finite order p and $\gamma_Z(h) = 0$ for $h \geq 1$, by Proposition 4.2.1 we have

$$\begin{aligned} \phi(B)Z(k) &= \phi(B)\pi(B)X(k) \\ &= \pi(B)\phi(B)X(k) \\ &= \pi(B)\theta(B)Z(k) \end{aligned} \tag{4.10}$$

Letting $\eta(z) = \pi(z)\theta(z) = \sum_{i=0}^{\infty} \eta_i z^i$ for $|z| \leq \sqrt{2}^{-1}$, we see from (4.10) that

$$\sum_{i=0}^{p} \frac{\phi_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k) = \sum_{i=0}^{\infty} \frac{\eta_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k).$$

Multiplying each side by $Z_{i'j'}(k)$ and taking expectations, we see that $\phi_i = \eta_i$ for $i = 0 \ldots p$, and $\eta_i = 0$ for $i > p$. With that, we have

$$\begin{aligned} \phi(z/\sqrt{2}) &= \eta(z/\sqrt{2}) \\ &= \pi(z/\sqrt{2})\theta(z/\sqrt{2}), \quad |z| < \sqrt{2}. \end{aligned}$$

Now, if $\theta(z/\sqrt{2}) = 0$ and $|\pi(z/\sqrt{2})| < \infty$ for $|z| \leq \sqrt{2}$, then $\phi(z) = 0$, hence $\phi(\cdot)$ and $\theta(\cdot)$ would have a common zero contradicting the assumption of no common zeros. Thus, $\theta(z/\sqrt{2}) \neq 0$ for any $|z| \leq \sqrt{2}$. ∎

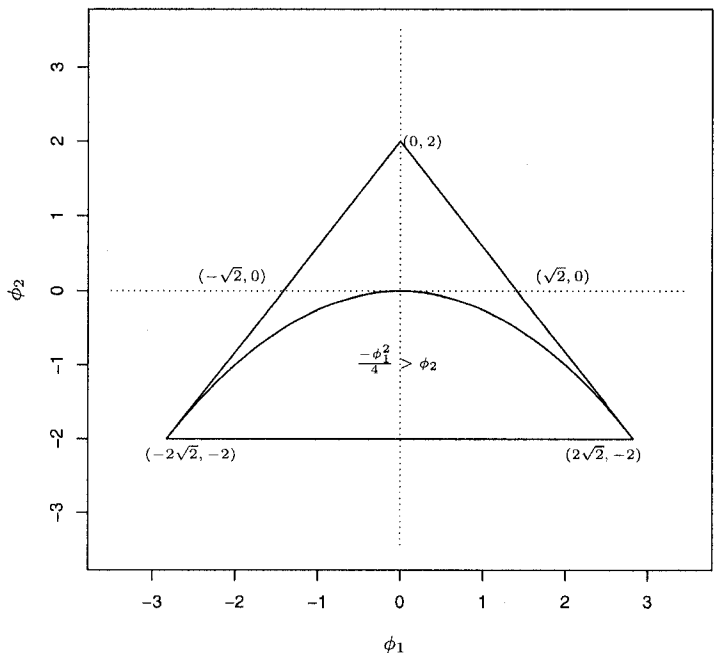


Figure 4.1: Region for causal AR(2)

Example 4.2.1 (The ARMA(1,1) Process)

The ARMA(1,1) defined by

$$X(k) - \frac{\phi}{2}(X(u_1) + X(u_2)) = Z(k) + \frac{\theta}{2}(Z(u_1) + Z(u_2)) \tag{4.11}$$

is both causal and invertible under the constraints that $|\phi| < \sqrt{2}$ and $|\theta| < \sqrt{2}$. The only root of $\phi(z/\sqrt{2})$ is $\sqrt{2}/\phi$, so if $\sqrt{2}/\phi > 1$, then $\phi < \sqrt{2}$ and $X(k)$ is causal. Similar arguments hold for determining the single constraint on θ.

Example 4.2.2 (The AR(2) Process)

The AR(2) defined by

$$(1 - \phi_1 B - \phi_2 B^2)X(k) = Z(k) \tag{4.12}$$

is causal when the all the roots of $\phi(z/\sqrt{2})$ are outside the unit circle. For an AR(2) to be causal, the parameters (ϕ_1, ϕ_2) must lie in the triangular region defined by the intersection of the three regions

$$\sqrt{2}\phi_1 + \phi_2 < 2,$$

$$\phi_2 - \sqrt{2}\phi_1 < 2,$$

$$|\phi_2| < 2$$

which can be seen in Figure 4.1. Causal autoregressive polynomials with complex roots are identified through the quadratic formula, which are easily seen to be those (ϕ_1, ϕ_2) in the causal region such that $\phi_2 < -\phi_1^2/4$.

4.3 The Autocovariance Function of a Causal/Invertible ARMA(p,q) Process

Only ARMA(p, q) processes that are both causal and invertible are considered. Two methods are presented for determining the autocovariance function (ACVF) for these processes.

4.3.1 Method I

By Theorem 4.2.1, we know that $\sum_{i=1}^{\infty} \frac{|\psi_i|}{\sqrt{2^i}} < \infty$. Since $Z(k) \sim WN(0, \sigma^2)$, then by Proposition 4.2.1, the ACVF $\gamma_X(k)$ is

$$\gamma_X(h) = \frac{\sigma^2}{2^h} \sum_{i=0}^{\infty} \frac{\psi_{i+h}\psi_i}{2^i}. \tag{4.13}$$

The coefficients ψ_i can be determined by matching coefficients for powers of B in the expression

$$(1 - \phi_1 B - \phi_2 B^2 \cdots - \phi_p B^p)(\psi_0 + \psi_1 B + \psi_2 B^2 + \ldots) = (1 + \theta_1 B + \theta_2 B^2 \cdots + \theta_q B^q)$$

where it is seen that

$$1 = \psi_0$$

$$\begin{aligned}
\theta_1 &= \psi_1 - \phi_1\psi_0 \\
\theta_2 &= \psi_2 - \phi_1\psi_1 - \phi_2\psi_0 \\
\theta_3 &= \psi_3 - \phi_1\psi_2 - \phi_2\psi_1 - \phi_3\psi_0 \\
&\vdots
\end{aligned}$$

or equivalently,

$$\psi_i - \sum_{j=1}^{p} \phi_j\psi_{i-j} = \theta_i, \quad i = 0, 1, \cdots \tag{4.14}$$

which can be solved recursively to obtain each ψ_i.

Another way to obtain ψ_i is via a solution to a system of equations with constant coefficients. For large i, (4.14) has the form of a p^{th} order difference equation. Using results in Brockwell and Davis (1991, Sec. 3.6), the general solution is written as

$$\psi_k = \sum_{i=1}^{s}\sum_{j=0}^{r_i-1} \alpha_{ij} k^j \xi_i^{-k}$$

where ξ_i are the s distinct roots of $\phi(z)$ and r_i is the multiplicity of ξ_i. Thus, $\sum_{i=1}^{s} r_i = p$, the order of $\phi(z)$. The α_{ij} and coefficients for $0 \leq j < \max(p, q+1) - p$ are determined by initial conditions.

Remark 4.3.1 The autocovariance function of a causal ARMA(p, q) is geometrically decreasing at a rate faster than $1/\sqrt{2}^h$ since

$$\begin{aligned}
\gamma_X(h) &< \frac{\sigma^2}{2^h}\left(2^h \sum_{i=0}^{\infty} \frac{\psi_{i+h}^2}{2^{i+h}}\right)^{\frac{1}{2}} \left(\sum_{i=0}^{\infty} \frac{\psi_i^2}{2^i}\right)^{\frac{1}{2}} \\
&< \frac{\sigma^2}{\sqrt{2}^h}\left(\sum_{i=0}^{\infty} \frac{\psi_i^2}{2^i}\right)
\end{aligned}$$

from which we see $\sum_{i=0}^{\infty} |\gamma_X(h)| < \infty$.

Example 4.3.1 For the ARMA(1,1), the coefficients ψ_k can be determined by matching coefficients

$$\psi_0 = 1$$

$$\begin{aligned}\psi_1 &= \phi + \theta \\ \psi_2 &= \phi\psi_1 \\ &\vdots \\ \psi_k &= \phi^{k-1}(\theta + \phi)\end{aligned}$$

which can then be used to determine the second moments. The alternative method using the general solution to a set of linear difference equations with constant coefficients and single root $\xi_1 = 1/\phi$,

$$\psi_k = a_{10}\phi^k, \quad k \geq 1$$

where a_{10} is determined by the initial conditions

$$\begin{aligned}\psi_0 &= 1 \\ \psi_1 - \psi_0\phi &= \theta.\end{aligned}$$

Using the general solution, we know $\psi_1 = a_{10}\phi = \theta + \phi$ from which it is easy to see that

$$a_{10} = \frac{\phi + \theta}{\phi}.$$

If $h = |k - k'|$, the variance is determined to be

$$\begin{aligned}\gamma(0) &= \sigma^2 \sum_{i=0}^{\infty} \frac{\psi_i^2}{2^i} \\ &= \sigma^2 \left(1 + \frac{(\phi+\theta)^2}{2} \sum_{i=0}^{\infty} \left(\frac{\phi^2}{2}\right)^j\right) \\ &= \sigma^2 \left(1 + \frac{(\phi+\theta)^2}{2(1-\phi^2/2)}\right) \qquad (4.15)\end{aligned}$$

when $\frac{\phi^2}{2} < 1$, a condition satisfied since $|\phi| < \sqrt{2}$ by the constraint on $\phi(z)$. Furthermore,

$$\gamma(1) = \frac{\sigma^2}{2}\left((\phi+\theta) + \frac{(\phi+\theta)^2\phi}{2(1-\phi^2/2)}\right) \qquad (4.16)$$

and

$$\gamma(h) = \left(\frac{\phi}{2}\right)^{h-1} \gamma(1) \tag{4.17}$$

for any $h \geq 2$.

The coefficients ψ_i are obviously not absolutely summable for any ϕ such that $1 < \phi < \sqrt{2}$. However, if $|\phi| < \sqrt{2}$, $\sum_{i=0}^{\infty} \frac{|\psi_i|}{\sqrt{2^i}}$ is still finite.

Remark 4.3.2 As with an MA(1) in time series, for any MA(1) with $\theta = (\theta, \sigma^2)$ and covariance $\gamma(h)$, there exists an alternative MA(1) representation with parameter vector $\theta' = (2/\theta, (\theta^2\sigma^2)/2)$ which has the same covariance. Since we only observe $Y(k)$, there is no way to distinguish between the two. One reasonable method for this is to set parameter constraints such that one is selected over the other. Constraints on θ that result in an invertible $Y(k)$ allow for this.

4.3.2 Method II

Recall that for any causal ARMA(p, q) process, $X(k)$ can be written as

$$X(k) = \sum_{i=0}^{\infty} \frac{\psi_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k).$$

Multiplying each side by any upstream $\mathbf{X}_{i,j}(k)$ and taking expectations, we find a system of equations defined by

$$\gamma(h) - \frac{\phi_1}{2} 2^{(1-h)_+} \gamma(h-1) - \cdots - \frac{\phi_p}{2^p} 2^{(p-h)_+} \gamma(|h-p|) = \sum_{j=h}^{q} \psi_{j-h} \frac{\theta_j}{2^j} \tag{4.18}$$

for $0 \leq h < m$ and

$$\gamma(h) - \frac{\phi_1}{2}\gamma(h-1) - \cdots - \frac{\phi_p}{2^p}\gamma(h-p) = 0 \tag{4.19}$$

for $h \geq m$ where $m = max(p, q+1)$ and $(x)_+ = \max\{0, x\}$. The coefficients ψ_{j-h} should be expressed in terms of elements in ϕ and θ through the expansion $\psi(z) = \theta(z)/\phi(z)$. Thus, for large h, we find $\gamma(h)$ to be in the form of a homogeneous linear

difference equation, where the solution for $\gamma(h)$ is a linear function of polynomial roots. Note however, that the roots are not those of $\phi(z)$, but those of $\phi(z/2)$. The p coefficients α_j as well as $\gamma(h), h = 0, \ldots, m-p$ can be obtained from the m equations defined by 4.18.

For a strictly AR(p) model, the above equations allow expression of the ACVF in terms of the polynomial roots of $\phi(z/2)$, where

$$\gamma(h) = \sum_{j=1}^{p} \alpha_j \xi_j^{-i}$$

can be directly substituted into (4.18) to obtain α_j. When $q > 0$, these equations are non-linear in θ, as will be seen in deriving initial conditions for optimization in the MA(1).

Example 4.3.2 For the strictly AR model of order 2, we have

$$\gamma(h) = \alpha_1 \xi_1^{-h} + \alpha_2 \xi_2^{-h},$$

which is substituted back into

$$\begin{aligned} \gamma(0) - \phi_1\gamma(1) - \phi_2\gamma(2) &= \sigma^2 \\ \gamma(1) - \frac{\phi_1}{2}\gamma(0) - \frac{\phi_2}{2}\gamma(1) &= 0 \end{aligned}$$

and then be solved to obtain expressions for α_1 and α_2. Furthermore, since $\phi(B)X(k) = (1 - \xi_1^{-1}B)(1 - \xi_2^{-1}B)X(k)$, we see that

$$\phi_1 = \xi_1^{-1} + \xi_2^{-1}$$

and

$$\phi_2 = -\xi_1^{-1}\xi_2^{-1}$$

where ξ_1 and ξ_2 are the roots of $\phi(z/2)$ (Recall that the solution to the homogeneous difference equations defined by (4.19) are in terms of the roots of $\phi(z/2)$ rather than $\phi(z/\sqrt{2})$). With substitution, we find

$$\alpha_i = \frac{\sigma^2 \xi_i^{-1}}{\left(2\xi_1^{-1}\xi_2^{-1} - 1\right)\left(1 - 2\xi_i^{-2}\right)\left(\xi_2^{-1} - \xi_1^{-1}\right)}$$

which is used to obtain the expression

$$\gamma(h) = \frac{\sigma^2 \xi_1^2 \xi_2^2}{\left(\xi_1^{-1}\xi_2^{-1} - 2\right)\left(\xi_2^{-1} - \xi_1^{-1}\right)} \left(\left(\xi_1^{-2} - 2\right)^{-1} \xi_1^{1-h} - \left(\xi_2^{-2} - 2\right)^{-1} \xi_2^{1-h} \right).$$

In the case of complex roots, we have $\xi_2 = \overline{\xi_1}$ and $\xi_1 = r\exp^{i\lambda} = r(cos(\lambda) + i\sin(\lambda)$ where

$$\lambda = \begin{cases} \tan^{-1}\left|\frac{\phi_1}{2\phi_2}\right| & \phi_1 < 0 \\ \pi - \tan^{-1}\left|\frac{\phi_1}{2\phi_2}\right| & \phi_1 > 0 \end{cases}.$$

Since $\phi_2 < 0$ for a causal AR(2) with complex roots, λ is restricted to $\lambda \in (0, \pi)$. With that, we derive the expression

$$\gamma(h) = \frac{\sigma^2 r^4 r^{-h} \left(r^2 \sin(\lambda + \lambda h) + 2\sin(\lambda - \lambda h)\right)}{\left(r^2 - 2\right)\left(r^4 - 4r^2\cos(2\lambda) + 4\right)\sin(\lambda)}.$$

4.3.3 Sample Autocovariance Function

Following Brockwell and Davis (1991, Chapter 7), define the sample autocovariance function to be

$$\hat{\gamma}(h) = n^{-1} \sum_{(k,k'):|k-k'|=h} \left(x(k) - \overline{x}\right)\left(x(k') - \overline{x}\right). \tag{4.20}$$

Note that the sum is over all occurrences of a particular lag h while the divisor is the total number of observations. In a time series context, using n as the divisor ensures a non-negative definite sample covariance matrix. Furthermore, for large n and reasonable h, the differences in $\hat{\gamma}(h)$ due the the different divisor is negligible. The impact of using the number of occurrences of lag h rather than n on the non-negative definiteness has not been determined for the tree structure here.

Once $\hat{\gamma}(h)$ is obtained, the sample autocorrelation function (ACF) easily follows by

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \tag{4.21}$$

The network analogue for a Partial Autocorrelation Function as well as the impact of a tree structure on the non-negative definite constraint of the estimated covariance matrix have not been considered, and are left as an area for future research.

4.4 State-Space Form

Any causal ARMA(p,q) model can be written in the state-space form given by (2.1) and (2.2). Furthermore, there may exist more than one state-space representation for a given causal ARMA(p,q) model. We present one such representation which is analogous to that for a time series (Brockwell and Davis, 1991, p.468), recognizing that others do exist, and may have system matrices with smaller dimension than those provided here.

The dimensions of the system matrices and vectors increase with p and q. If we define $r = \max(p, q+1)$, we can obtain a state-space representation for the general ARMA(p,q) model. Further define $\theta_0 = 1$ and note that $\phi_i = 0$ for $i > p$ whereas $\theta_i = 0$ for $i > q$. From (2.1), we find the observation equation for the ARMA(p,q) model to be

$$\mathbf{Y}(k) = \left[\begin{array}{cccc} \frac{\theta_{r-1}}{2^{r-1}}^T & \frac{\theta_{r-2}}{2^{r-2}}^T & \cdots & \theta_0 \end{array} \right] \mathbf{X}(k)$$

where $\left[\theta_j/(2^j)\right]^T$ is a 1×2^j vector with each element equal to $\theta_j/(2^j)$. The state vector is defined to be

$$\mathbf{X}(k) = \left[\begin{array}{c} \mathbf{X}_{r,1}(k) \\ \mathbf{X}_{r,2}(k) \\ \mathbf{X}_{r-1,1}(k) \\ \mathbf{X}_{r-1,2}(k) \\ \vdots \\ X_{1,1}(k) \\ X_{1,2}(k) \\ X(k) \end{array} \right]$$

where $\mathbf{X}_{h,j}(k)^T$ is a vector of elements $X(k')$ such that $|k-k'| = h$ and k' is upstream of parent j. In the state equation, we see a block-like structure since $\mathbf{X}(k)$ consists of elements from each parent. The system matrix associated with parent i is defined by

$$F_{i,k} = \left[\begin{array}{cccc} & & F^*_{i,k} & \\ \frac{\phi_r}{2^r} & \frac{\phi_{r-1}}{2^{r-1}} & \cdots & \frac{\phi}{2} \end{array} \right]$$

where the rows of $F^*_{i,k}$ are defined to select the appropriate elements from $\mathbf{X}(u_i)$.

Example 4.4.1 The ARMA(2,2)

In a state-space representation, the observation equation for the univariate ARMA(2,2) can be written as

$$Y(k) = \left[\begin{array}{cccccccc} \theta_2/4 & \theta_2/4 & \theta_2/4 & \theta_2/4 & \theta_1/2 & \theta_1/2 & 1 \end{array} \right] \mathbf{X}(k) \tag{4.22}$$

where

$$\mathbf{X}(k) = \left[\begin{array}{c} X_{2,1}(k) \\ X_{2,2}(k) \\ X_{2,3}(k) \\ X_{2,4}(k) \\ X_{1,1}(k) \\ X_{1,2}(k) \\ X(k) \end{array} \right].$$

The corresponding matrix in the state equation associated with parent u_1 is defined by

$$F_{1,k} = \left[\begin{array}{ccccccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \phi_2/4 & \phi_2/4 & \phi_1/2 \end{array} \right] \tag{4.23}$$

where the matrix for parent u_2 similarly follows. With that, the state-space representation for the ARMA$(2,2)$ is easily obtained. The covariance matrix associated with the white noise process is defined by

$$Q(k) = \left[\begin{array}{cc} 0_{6\times 6} & \mathbf{0}_{6\times 1} \\ \mathbf{0}_{1\times 6} & \sigma^2 \end{array} \right].$$

Parameter constraints for the stationary case are established through the concepts of causality and invertibility of a process on a stream network.

4.5 q-Correlated Processes

Here we present the network analogue to the existence of an MA(q) representation for a q-correlated process in time series. If h denotes a previous generation

of $X(k)$, then for a q-correlated process, $X(k)$ is correlated to every $X_{h,j}(k)$, for $h = 1, \ldots, q$ but uncorrelated with any $X_{h,j}$ for $h > q$. We further assume the correlation is constant with every reach that exists at lag h, but changes with h. We restrict ourselves to binary trees.

Proposition 4.5.1 *Suppose $\{X(k)\}$ is a zero mean stationary process on a network such $\gamma(h) = Cov(X(k), X_{h,j}(k)) = 0$ for $h > q$ but $\gamma(q) \neq 0$. Then there exists a white noise sequence $\{Z(k)\}$ such that*

$$X(k) = Z(k) + \sum_{i=1}^{q} \frac{\theta_i}{2^i} \sum_{j=1}^{2^i} Z_{ij}(k),$$

so that $\{X(k)\}$ is an MA(q) process.

Proof The proof parallels that of the moving average representation for a q-correlated time series in Brockwell and Davis (1991, p. 90). Results of the Projection Theorem (Brockwell and Davis, 1991, p. 51) are applied to the tree structure of a stream network. If $\mathcal{M}$ is a closed subspace of the Hilbert space $\mathcal{H}$, then for each $x \in \mathcal{H}$ there is a unique $\hat{x} \in \mathcal{M}$ such that $x - \hat{x} \in \mathcal{M}^{\perp}$, the orthogonal complement of $\mathcal{M}$. The required mapping is $\mathcal{P}_{\mathcal{M}} x = \hat{x}$. The subspaces considered here are closed spans of random variables from the Hilbert space $L^2(\Omega, \mathcal{F}, P)$.

Let $\mathcal{X}_{(k)} = \mathcal{X}_{(k]} \backslash \{X(k)\}$ where $\mathcal{X}_{(k]} = \mathbf{X}(k) \cup \{\mathbf{X}_{i,j}(k)\}_{i=1\ldots\infty, j=1\ldots 2^i}$. Then for each reach k, define the subspaces $\mathcal{M}_{(k)} = \overline{sp}\{\mathcal{X}_{(k)}\}$ and $\mathcal{M}_{(k]} = \overline{sp}\{\mathcal{X}_{(k]}\}$, and let

$$Z(k) \;=\; X(k) - \mathcal{P}_{\mathcal{M}_{(k)}} X(k)$$

from which $Z(k) \in \mathcal{M}_{(k]}$, yet $Z(k) \in \mathcal{M}_{(k)}^{\perp}$, a direct result of the Projection Theorem. Thus if k' is upstream of k, then $Z(k') \in \mathcal{M}_{(k']} \subset \mathcal{M}_{(k)}$ and hence $E\left(Z(k')Z(k)\right) = 0$. To show that $Z(k)$ is stationary, a convergence result of projection mappings is needed. Specifically, for $\mathcal{X}_{n,(k)} = \{X_{i,j}(k)\}_{i=1\ldots n, j=1\ldots 2^i}$,

$$\mathcal{P}_{\overline{sp}\{\mathcal{X}_{n,(k)}\}} X(k) \xrightarrow{m.s.} \mathcal{P}_{\mathcal{M}_{(k)}} X(k)$$

as $n \longrightarrow \infty$. To show this, define the normalized value of $X(k)$ to be

$$I(k) = \frac{X(k) - \mathcal{P}_{\mathcal{M}_{(k)}} X(k)}{\left\| X(k) - \mathcal{P}_{\mathcal{M}_{(k)}} X(k) \right\|}$$

so that $I(k)$ are linear combinations of elements in $\mathcal{X}_{(k]}$ implying that $\overline{sp}\{\mathcal{X}_{(k]}\} = \overline{sp}\{\mathcal{I}_{(k]}\}$ for $\mathcal{I}_{(k]} = I(k) \cup \{I_{i,j}(k)\}_{i=1...\infty, j=1...2^i}$. Since $\overline{sp}\{\mathcal{I}_{(k)}\}$ is a separable Hilbert space, the desired result follows from Brockwell and Davis (1991, Section 2.4).

By stationarity of $X(k)$ and the continuity of the L^2 norm,

$$\begin{aligned}
\|Z(k)\| &= \left\| X(k) - \mathcal{P}_{\mathcal{M}_{(k)}} X(k) \right\| \\
&= \lim_{n \to \infty} \left\| X(k) - \mathcal{P}_{\overline{sp}\{\mathcal{I}_{n,(k)}\}} X(k) \right\| \\
&= \lim_{n \to \infty} \left\| X(u_1) - \mathcal{P}_{\overline{sp}\{\mathcal{I}_{n,(u_1)}\}} X(u_1) \right\| \\
&= \left\| X(u_1) - \mathcal{P}_{\mathcal{M}_{(u_1)}} X(u_1) \right\| \\
&= \|Z(u_1)\|
\end{aligned}$$

Defining $\|Z(k)\|^2 = \sigma^2$, it is clear that $Z(k) \sim \mathrm{WN}(0, \sigma^2)$. To continue, we seek to obtain a decomposition of $\mathcal{M}_{(k)}$ into orthogonal subspaces. We see that

$$\begin{aligned}
\mathcal{M}_{(k)} &= \overline{sp}\{\mathcal{X}_{(u_1]}, \mathcal{X}_{(u_2]}\} \\
&= \overline{sp}\{\mathcal{X}_{(u_1)}, \mathcal{X}_{(u_2)}, Z(u_1), Z(u_2)\} \\
&\vdots \\
&= \overline{sp}\{\cup \mathcal{X}_{(k')}, |k - k'| = q + 1, Z_{i,j}(k), i = 1...q, j = 1...2^i\}
\end{aligned}$$

Since $X(k)$ is q-correlated, it follows that $X(k) \perp \overline{sp}\{\cup \mathcal{X}_{(k')}, |k - k'| = q + 1\}$. Noting that $\mathcal{M}_{(k)}$ is itself a Hilbert space which contains $\mathcal{P}_{\mathcal{M}_{(k)}} X(k)$ and that $\overline{sp}\{Z_{i,j}(k), i = 1...q, j = 1...2^i\} \subset \mathcal{M}_{(k)}$ then using properties of projection mappings and orthonormal spaces,

$$\begin{aligned}
\mathcal{P}_{\mathcal{M}_{(k)}} X(k) &= \mathcal{P}_{\overline{sp}\{\cup \mathcal{X}_{(k')}, |k-k'|=q+1\}} \mathcal{P}_{\mathcal{M}_{(k)}} X(k) + \mathcal{P}_{\overline{sp}\{Z_{i,j}(k), i=1...q\}} \mathcal{P}_{\mathcal{M}_{(k)}} X(k) \\
&= 0 + \mathcal{P}_{\overline{sp}\{Z_{i,j}(k), i=1...q\}} X(k)
\end{aligned}$$

$$= \sum_{i=1}^{q} \sum_{j=1}^{2^i} \left\langle X(k), \frac{Z_{i,j}(k)}{\sigma} \right\rangle \frac{Z_{i,j}(k)}{\sigma}$$
$$= \sum_{i=1}^{q} \frac{E\left(X(k)Z_{i,j}(k)\right)}{\sigma^2} \sum_{j=1}^{2^i} Z_{i,j}(k)$$

since we assume that $E\left(X(k)Z_{i,j}(k)\right)$ is constant for all j at a fixed lag i and is independent of k by stationarity. If we define

$$\frac{\theta_i}{2^i} = \frac{E\left(X(k)Z_{i,j}(k)\right)}{\sigma^2},$$

then we have

$$\begin{aligned} Z(k) &= X(k) - \mathcal{P}_{\mathcal{M}_{(k)}} X(k) \\ &= X(k) - \sum_{i=1}^{q} \frac{\theta_i}{2^i} \sum_{j=1}^{2^i} Z_{i,j}(k) \end{aligned}$$

from which the Moving Average representation is immediate.

4.6 Simulation Results

The AR(2) as well as the ARMA(1,1) with special cases AR(1) and MA(1) are considered on two different tree structures. The first is a full binary tree with 8 levels, which consists of 255 observations. Since there only exists 8 levels in this binary tree, the largest lag possible is 7. The second tree is that defined by a subset of Upper and Lower Rock Creek in Montgomery County, Maryland, as seen in Figure 4.2. This network structure consists of two running second order segments that merge to create a third order segment. Each of these higher order segments consists of several inputs of first order reaches. There are 39 reaches in total, with lags 0 through 15 present, where there are very few occurrences of these higher lags.

Realizations were generated for each of these lower order ARMA models with several different parameter values. For all models, we defined the true variance of the

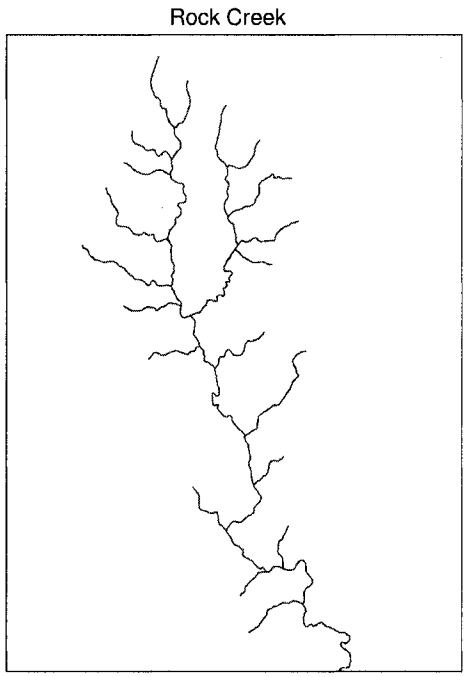


Figure 4.2: Portion of Upper and Lower Rock Creek, Montogmery County, Maryland, consisting of 21 first order reaches, 10 second order reaches, and 9 third order reaches. There are 2 second order segments merging to form 1 third order segment.

white noise process to be $\sigma^2 = 1$. Values for the autoregressive parameter are $\phi \in \{-1.2, -0.6, 0.6, 1.2\}$ whereas those considered for the moving average parameter are $\theta \in \{-1.1, -0.4, 0.4, 1.1\}$. Note that different values for ϕ and θ are required since the stationary ARMA(p, q) models assume polynomials with no common zeros.

Realizations for the AR(2) were obtained with model parameters chosen from the causal region (defined in Figure 4.1) that represent several different forms of dependence.

4.6.1 Data Generation

4.6.1.1 ARMA(1,1)

For each reach k, values for $Y(k)$ and $Z(k)$ are obtained and made available for the immediate downstream recursion. For the first order reaches, we see that

$$\begin{bmatrix} Y(k) \\ Z(k) \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2\left(1+\frac{(\phi+\theta)^2}{2(1-\phi^2/2)}\right) & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix}\right)$$

from which $i.i.d$ draws generate the simulated $[Y(k), Z(k)]$. At each higher order reach, white noise was drawn from $N(0, \sigma^2)$, which was then used with the downstream recursive relationship to obtain $Y(k)$.

4.6.1.2 AR(2)

The covariance $\gamma(h), h = 0, 1$ was determined for specified (ϕ_1, ϕ_2) in the region defined in Figure 4.1 using (4.18) and (4.19). Then for each first order reach, independent draws were obtained from

$$\begin{bmatrix} Y(u_1) \\ Y(u_2) \\ Y(k) \\ Z(k) \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \gamma(0) & 0 & \phi_1/2\gamma(1) & 0 \\ 0 & \gamma(0) & \phi_1/2\gamma(1) & 0 \\ \phi_1/2\gamma(1) & \phi_1/2\gamma(1) & \gamma(0) & \sigma^2 \\ 0 & 0 & \sigma^2 & \sigma^2 \end{bmatrix}\right)$$

where the values for the imaginary parents are needed for the recursive relationship with the reaches immediately downstream of first order k. Values $Y(k)$ were easily obtained for all higher order reaches through random draws of independent noise and the recursive relationship as defined by the model.

4.6.2 Kalman Recursions

4.6.2.1 ARMA(1,1)

The state-space form for the ARMA(1,1) follows from §4.4. Here we define $\mathbf{X}(k)^T = \begin{bmatrix} X(u_1) & X(u_2) & X(k) \end{bmatrix}^T$. A state-space representation for the ARMA(1,1) is defined by

$$Y(k) = \begin{bmatrix} \theta/2 & \theta/2 & 1 \end{bmatrix} \mathbf{X}(k)$$

$$\mathbf{X}(k) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & \phi/2 \end{bmatrix} \mathbf{X}(u_1) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \phi/2 \end{bmatrix} \mathbf{X}(u_2) + \begin{bmatrix} 0 \\ 0 \\ Z(k) \end{bmatrix}$$

where

$$Q(k) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

for all k.

Because of stationarity, initial conditions for the state at each first order reach are easily identified. The initial predictions and corresponding variances are defined by the model assumptions where $X(k)$ has mean zero with $\text{Var}[X(k)] = \frac{\sigma^2}{1-\phi^2/2}$. If we assume that each first order reach has two imaginary upstream parents, we have

$$\mathbf{X}^p(k) = \mathbf{0}, \quad \Omega_k^p = \frac{\sigma^2}{1-\phi^2/2} \begin{bmatrix} 1 & 0 & \phi/2 \\ 0 & 1 & \phi/2 \\ \phi/2 & \phi/2 & 1 \end{bmatrix},$$

which are used to initialize the filter. The innovation variance Δ_k is derived to be

$$\Delta_k = \frac{\sigma^2}{1-\phi^2/2}\left(1 + \phi\theta + \frac{\theta^2}{2}\right).$$

4.6.2.2 AR(2)

The state-space form of an AR(2) is defined by setting

$$G_k = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

with

$$F_{k,u_1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ \phi_2/4 & \phi_2/4 & \phi_1/2 \end{bmatrix} \quad F_{k,u_2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ \phi_2/4 & \phi_2/4 & \phi_1/2 \end{bmatrix}$$

and

$$Q(k) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}.$$

As with the ARMA(1,1), $W(k)$ is defined to be zero for all k. Initial conditions for the states at first order reaches are defined by the unconditional distribution of $X(k)$, which is the same for all first order k by flow-connectedness and stationarity. For a specified (ϕ_1, ϕ_2) in the region defined in Figure 4.1, the covariance structure is obtained through either of the methods discussed in §4.3. With that, we have

$$\mathbf{X}^p(k) = \mathbf{0}, \quad \Omega_k^p = \begin{bmatrix} \gamma(0) & 0 & \gamma(1)\phi_1/2 \\ 0 & \gamma(0) & \gamma(1)\phi_1/2 \\ \gamma(1)\phi_1/2 & \gamma(1)\phi_1/2 & \gamma(0) \end{bmatrix}$$

assuming each first order reach has two imaginary upstream parents simply for convenience. The innovation variance Δ_k follows immediately from

$$\Delta_k = G_k \Omega_k^p G_k^T$$

since $\mathbf{W}(k) = 0$ for all k.

4.6.3 Gaussian Likelihood

The likelihood in terms of the innovations is easily obtained through execution of the Kalman recursions. Since neither the prediction nor innovation variances depend on the data, we can see that

$$\Omega_k^p = \sigma^2 \Omega_k^{*p} \quad \Delta_k = \sigma^2 \Delta_k^*$$

where Ω_k^{*p} and Δ_k^* can both be obtained from the usual recursions with $\sigma^2 = 1$. Furthermore, the innovations $\mathbf{v}(k)$ do not depend on σ^2. With this, we see that apart from a constant, -2 times the log of the likelihood is

$$l(\mathbf{Y}, \sigma^2, \phi, \theta) = \sum_{i=1}^{m_s} \sum_{j \in \mathcal{S}_i} \sum_{k=1}^{n_{ij}} \left(\log \sigma^2 + \log |\Delta_k^*| + \frac{1}{\sigma^2} \mathbf{v}(k)^T \Delta_k^{*-1} \mathbf{v}(k) \right)$$

which can be used to obtain the closed from expression for the maximum likelihood estimator

$$\hat{\sigma^2} = \frac{1}{N} \sum_{i=1}^{m_s} \sum_{j \in \mathcal{S}_i} \sum_{k=1}^{n_{ij}} \mathbf{v}(k)^T \Delta_k^{*-1} \mathbf{v}(k)$$

where N is the total number of reaches. With substitution, we have a concentrated likelihood

$$l_c(\mathbf{Y}, \phi, \theta) = N \log \hat{\sigma}^2 + \sum_{i=1}^{m_s} \sum_{j \in \mathcal{S}_i} \sum_{k=1}^{n_{ij}} \log |\Delta_k^*| + N$$

which is now a function in fewer parameters.

4.6.4 Parameter Estimates

A network analogue of Yule-Walker type estimators can be obtained by using (4.18) and (4.19). For a strictly AR(p) model, we obtain the Yule-Walker equations

$$\Gamma_p \phi = \gamma_p \tag{4.24}$$

and

$$\sigma^2 = \gamma(0) - \phi^T \gamma_p \tag{4.25}$$

where Γ_p is the matrix $\left[\frac{\gamma(i-j)2^{(j-i)_+}}{2^j}\right]_{i,j=1}^{p}$ for row i, column j, and $\gamma_p = (\gamma(1), ..., \gamma(p))$. Note that Γ_p is similar in form to that in time series, with the difference being specific weights for each element because of the upstream averaging.

If we replace the covariances $\gamma(h), h = 0, ..., p$ with their corresponding sample covariances $\hat{\gamma}(h)$, we obtain the estimates

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p \tag{4.26}$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}^T \hat{\gamma}_p \tag{4.27}$$

where $\hat{\Gamma}_p$ and $\hat{\gamma}_p$ are defined as was for (4.24) and (4.25). Unlike the case in time series, the estimate defined by (4.26) does not guarantee that $1 - \hat{\phi}_1 z - \cdots - \hat{\phi}_p z^p \neq 0$ for $z \leq \sqrt{2}^{-1}$ (Brockwell and Davis, 1991, p.240).

Using (4.18) and (4.19), the Yule-Walker equations can be generalized for cases when $q > 0$. If we express ψ_{j-i} in terms of elements in ϕ and θ through the expansion $\psi(z) = \theta(z)/\phi(z)$, a system of equations is easily obtained. However, as we will see with the MA(1), these equations can be non-linear in the unknown model parameters.

Maximum likelihood was another tool which was used to obtain estimates of model parameters. We have shown a closed form expression for $\hat{\sigma^2}$, and numerical optimization routines can be used to obtain estimates for the remaining ϕ and θ parameters. Moreover, estimators that are solutions to the Yule-Walker equations can be used as preliminary estimates required for these numerical techniques. Other estimators from Burg's algorithm, the Innovations Algorithm, and Hannan-Rissanen Algorithm may be adapted to tree structures, but these were not considered here.

In order to obtain estimators which result in a causal/invertible ARMA process, optimization needs to consider the necessary parameter constraints. When applicable, a transformation was used in an effort to minimize numerical difficulties because of these boundary conditions. The optimization in the following exercise was chosen to be an unconstrained minimization over ξ_{new} through the transform

$$\xi_{new} = \sqrt{\frac{\xi_{old}^2}{a^2 - \xi_{old}^2}}$$

for $\xi_{old} \in (-a, a)$ as suggested by Durbin and Koopman (2001, p.142). Using this transformation, iterative estimates were less sensitive to the boundary conditions. For models such that a simple transformation was not available such as the AR(2), other techniques were considered.

The optimization routine used for maximization requires initial starting values. To assure that the estimate is truly one that maximizes the likelihood, four starting points within a causal/invertible region were also used to initialize the routine. From the different starting points for which the optimizations converged, we selected those which resulted in the highest likelihood.

4.6.5 First Order Autoregressive: AR(1)

Realizations for $\phi \in \{-1.2, -0.6, 0.6, 1.2\}$ were generated, where $\sigma^2 = 1$. The state vectors for the AR(1) can be 1-dimensional for simplicity. In this case, $Y(k) = X(k)$ and the Kalman Filter is initialized with

$$X^p(k) = 0 \quad \Omega_k^p = \frac{\sigma^2}{1-\phi^2/2}$$

for any first order reach k. Clearly, $\Delta_k = \Omega_k^p$ for these first order reaches. Since the prediction variances do not depend on the data, we see that $\Delta_k = \Omega_k^p = \sigma^2$ for any downstream reach with observed data. This allows even further simplification of the log-likelihood, seen by

$$\begin{aligned} l(\mathbf{Y}, \sigma^2, \phi) \quad \propto \quad & \frac{-N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{k \in \mathcal{H}} \left(Y(k) - \frac{\phi}{2}(Y(u_1) + Y(u_2)) \right)^2 \\ & + \frac{n_1}{2} \log(1 - \phi^2/2) - \frac{1}{2\sigma^2} \sum_{k \in \mathcal{F}} Y(k)(1 - \phi^2/2) \end{aligned}$$

where n_1 is the number of first order reaches. In this function, the variance σ^2 is easily concentrated out, and the $\log(1 - \phi^2/2)$ terms control its behavior near the boundaries of $|\phi| = \sqrt{2}$.

For each realization, maximum likelihood estimates for ϕ and σ^2 were obtained from which estimates for $\gamma(h)$ and $\rho(h)$ are calculated. Using the Yule-Walker equation (4.26), a method of moments estimator was generated by

$$\hat{\phi} = \frac{2\hat{\gamma}(1)}{\hat{\gamma}(0)}$$

where $\hat{\gamma}(h)$ was obtained from the sample ACVF using (4.20). From this we obtain the Yule-Walker estimate by

$$\hat{\phi}_{YW} = \begin{cases} \hat{\phi} & |\hat{\phi}| \leq \sqrt{2} \\ \text{sign}(\hat{\phi})\sqrt{2} & |\hat{\phi}| > \sqrt{2} \end{cases}$$

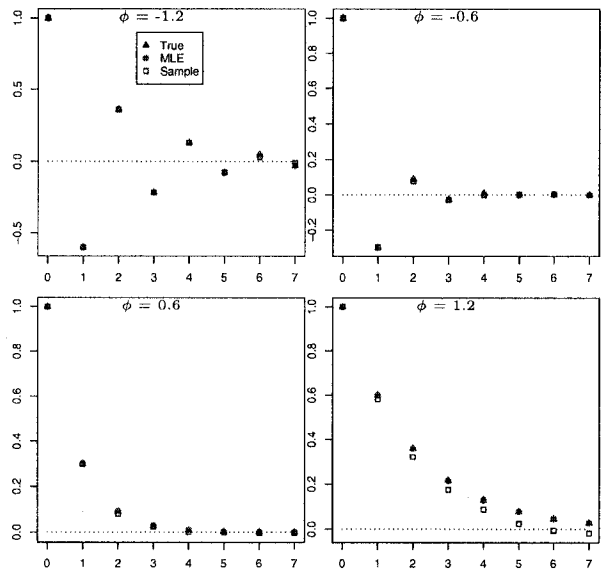
where we set the estimate to be the boundary if the moment estimate represented a non-causal process. The four starting points in the numerical optimization corresponded to $\phi \in \{-1, -.3, .3, 1\}$.

The resulting ACF from averaging the sample ACF over the simulated realizations can be seen in Figure 4.3. Parameter estimates from the 100 realizations were also averaged. It is easy to see from Figure 4.3 that the correlation decays exponentially. Furthermore, we see that the sample ACF tends to under-estimate the truth, especially with the Rock Creek structure where there are relatively few occurrences of each lag available for estimation.

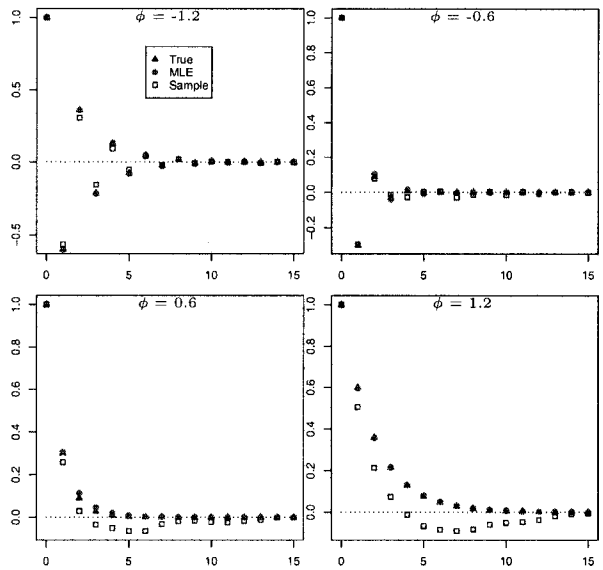
For models where $|\phi| = 1.2$, the optimization often resulted in estimates close to the boundaries of $\pm\sqrt{2}$, clearly indicated by the skewness seen in Figure 4.4. When this occurred, the estimate for σ^2 tended to be larger, and consequently that for $\gamma(0)$ was greatly inflated. Results for the Maximum Likelihood estimators are found in Tables 4.1 and 4.2, whereas those for Yule-Walker estimators are in Tables 4.3 and 4.4.

In Tables 4.1 and 4.2, we see that mean square errors associated with each estimate are much smaller for the binary tree. This is expected since the number of reaches is more than six times that on Rock Creek. There is also much less variation in the estimates obtained via maximum likelihood.

We also see from Tables 4.3 and 4.4 that unlike the case in time series, the moment estimators resulting from the Yule-Walker equations are not guaranteed to represent a causal model. Moreover, the associated variance estimate is not guaranteed to be non-negative. This situation seems to be more prominent on the Rock Creek structure. We also note a larger bias for $\hat{\sigma^2}$ when the true ϕ is near the boundary. Exploratory plots indicate that the downward biased $\hat{\phi}$ (ie, biased towards zero) results in positively biased estimate of the white noise variance. Because of this inverse relation between estimators, the realizations for which $|\hat{\phi}| > \sqrt{2}$ resulted in a negative variance estimate.



(a) Binary tree



(b) Rock Creek structure.

Figure 4.3: True, estimated (MLE), and empirical ACF for AR(1) processes on different tree structures. The MLE and empirical ACF displayed result from averaging the corresponding function over the 100 simulated realizations.

Table 4.1: Maximum Likelihood Estimates from 100 simulated realizations of each AR(1) process on a Binary Tree

ϕ	$E(\hat{\phi})$	$Bias(\hat{\phi})$	$MSE(\hat{\phi})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
-1.2	-1.2	0	0.001	1	0.983	-0.017	0.013	0
-0.6	-0.59	0.01	0.007	1	0.979	-0.021	0.008	0
0.6	0.6	0	0.01	1	0.985	-0.015	0.009	0
1.2	1.196	-0.004	0.001	1	1.011	0.011	0.014	0

[1] Number of realizations (out of 100) with non-convergent optimization.

Table 4.2: Maximum Likelihood Estimates from 100 simulated realizations of each AR(1) process on Rock Creek

ϕ	$E(\hat{\phi})$	$Bias(\hat{\phi})$	$MSE(\hat{\phi})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
-1.2	-1.181	0.019	0.01	1	0.947	-0.053	0.097	0
-0.6	-0.566	0.034	0.075	1	0.927	-0.073	0.053	0
0.6	0.573	-0.027	0.086	1	0.977	-0.023	0.052	0
1.2	1.173	-0.027	0.013	1	0.994	-0.006	0.09	0

[1] Number of realizations (out of 100) with non-convergent optimization.

Table 4.3: Yule-Walker Estimates from 100 simulated realizations of each AR(1) process on a Binary Tree

ϕ	$E(\hat{\phi})$	$Bias(\hat{\phi})$	$MSE(\hat{\phi})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
-1.2	-1.187	0.013	0.013	1	0.99	-0.01	0.175	2
-0.6	-0.59	0.01	0.011	1	0.974	-0.026	0.009	0
0.6	0.589	-0.011	0.015	1	0.983	-0.017	0.01	0
1.2	1.15	-0.05	0.016	1	1.13	0.13	0.208	3

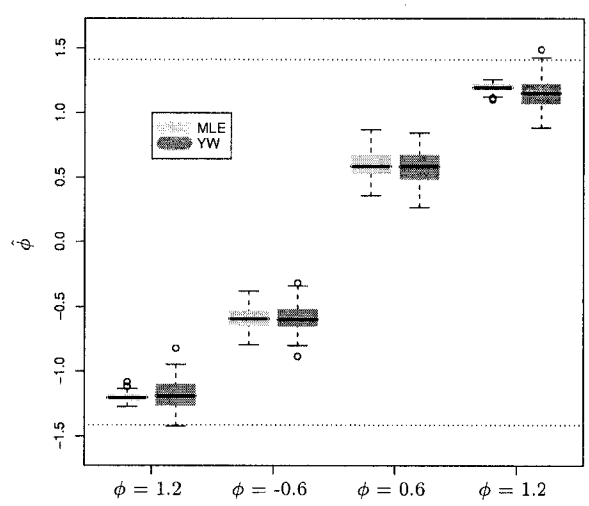
[1] Number of realizations (out of 100) with non-causal YW estimates.

Table 4.4: Yule-Walker Estimates from 100 simulated realizations of each AR(1) process on Rock Creek

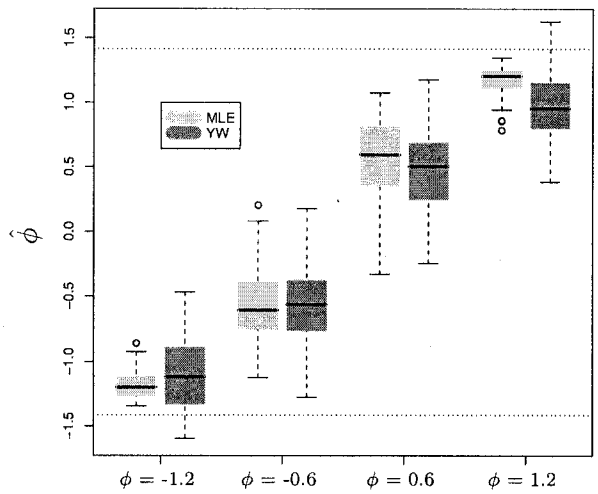
ϕ	$E(\hat{\phi})$	$Bias(\hat{\phi})$	$MSE(\hat{\phi})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
-1.2	-1.08	0.12	0.086	1	1.104	0.104	0.712	15
-0.6	-0.578	0.022	0.088	1	0.9	-0.1	0.079	0
0.6	0.476	-0.124	0.116	1	0.983	-0.017	0.063	0
1.2	0.969	-0.231	0.107	1	1.48	0.48	0.846	5

[1] Number of realizations (out of 100) with non-causal YW estimates.

The tabular results for the Maximum Likelihood and Yule-Walker estimates are graphically summarized in Figure 4.4.



(a) Binary Tree



(b) Rock Creek structure.

Figure 4.4: Maximum Likelihood (MLE) and Yule-Walker Method-of-Moments (YW) parameter estimates for 100 simulated realizations of each AR(1) process on different tree structures. Note that YW estimates are defined to be the boundary value when the moment estimates are outside the causal region.

4.6.6 First Order Moving Average: MA(1)

Realizations for $\theta \in \{-1.1, -0.4, 0.4, 1.1\}$ were generated, where $\sigma^2 = 1$. As with the AR(1), a sample ACVF and ACF are calculated using (4.20) and (4.21). The resulting ACFs from averaging the sample ACF and MLE ACFs over the 100 simulated realizations can be seen in Figure 4.5. A method of moments estimator was determined from the relation

$$\begin{aligned} \gamma(0) &= \sigma^2\left(1+\frac{\theta^2}{2}\right) \\ \gamma(1) &= \sigma^2\frac{\theta}{2} \end{aligned}$$

which are non-linear in θ. The solution

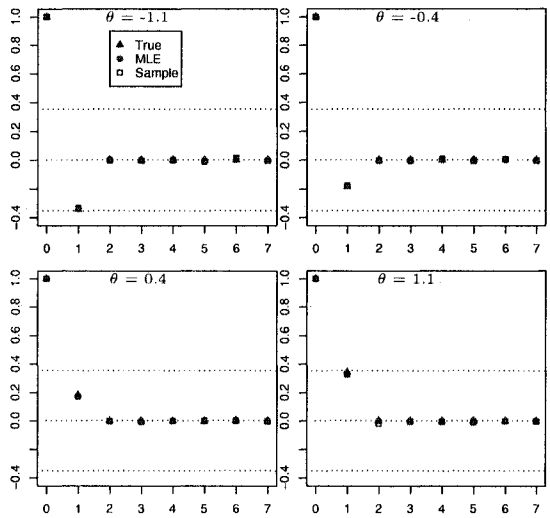
$$\theta = \frac{1-\sqrt{1-8\rho(1)^2}}{2\rho(1)}$$

results in an invertible process and is real-valued when $\rho \leq 1/2\sqrt{2}$. In the event that $|\hat{\rho}(1)| > 1/2\sqrt{2}$, the moment estimate for θ was defined to be the closest boundary. The numbers of realizations for which this did not occur are summarized in the last columns Table 4.7 and 4.8. The moment estimator for θ is then defined by

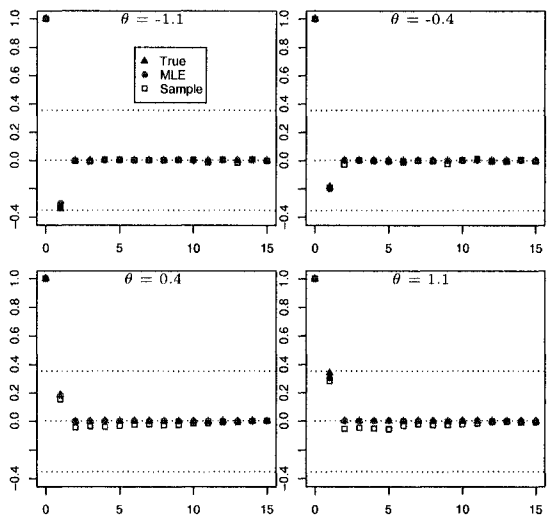
$$\hat{\theta} = \begin{cases} \frac{1-\sqrt{1-8\hat{\rho}(1)^2}}{2\rho(1)} & |\hat{\rho}(1)| \leq \frac{1}{2\sqrt{2}} \\ \text{sign}(\hat{\rho}(1))\sqrt{2} & |\hat{\rho}(1)| > \frac{1}{2\sqrt{2}} \end{cases}.$$

Maximum likelihood estimates for θ and σ^2 were also determined, with initial values corresponding to $\theta \in \{-1, -.3, .3, 1\}$. The estimate selected for summarization was the one that resulting in the highest likelihood. Results from Maximum Likelihood and Moment estimation for each of these 100 realizations are summarized in Tables 4.5, 4.6, 4.7 and 4.8.

From Figure 4.5, it is obvious that both the maximum likelihood and sample estimates of the autocorrelation are approximately zero after lag 1. It is also noticed that the true correlation at lag 1 is bounded by $1/2\sqrt{2} \approx 0.35$, indicated by the horizontal lines above and below zero, a necessity for an invertible MA(1).



(a) Binary Tree



(b) Rock Creek structure

Figure 4.5: True, estimated (MLE), and empirical ACF for MA(1) processes on different tree structures. The MLE and empirical ACF displayed result from averaging the corresponding function over the 100 simulated realizations.

Table 4.5: Maximum Likelihood Estimates from 100 simulated realizations of each MA(1) process on a Binary Tree

θ	$E(\hat{\theta})$	$Bias(\hat{\theta})$	$MSE(\hat{\theta})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
-1.1	-1.133	-0.033	0.058	1	0.971	-0.029	0.02	0
-0.4	-0.393	0.007	0.015	1	0.987	-0.013	0.007	0
0.4	0.386	-0.014	0.019	1	0.999	-0.001	0.006	0
1.1	1.127	0.027	0.054	1	0.971	-0.029	0.027	0

[1] Number of realizations (out of 100) with non-convergence in Maximum Likelihood estimation.

Table 4.6: Maximum Likelihood Estimates from 100 simulated realizations of each MA(1) process on Rock Creek

θ	$E(\hat{\theta})$	$Bias(\hat{\theta})$	$MSE(\hat{\theta})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
-1.1	-1.063	0.037	0.191	1	0.951	-0.049	0.098	0
-0.4	-0.543	-0.143	0.225	1	0.893	-0.107	0.066	0
0.4	0.464	0.064	0.217	1	0.939	-0.061	0.065	0
1.1	1.02	-0.08	0.201	1	1.015	0.015	0.083	0

[1] Number of realizations (out of 100) with non-convergence in Maximum Likelihood estimation.

Table 4.7: Method of Moments Estimates from 100 simulated realizations of each MA(1) process on a Binary Tree

θ	$E(\hat{\theta})$	$Bias(\hat{\theta})$	$MSE(\hat{\theta})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NI [1]
-1.1	-1.029	0.071	0.109	1	1.081	0.081	0.042	32
-0.4	-0.393	0.007	0.02	1	0.983	-0.017	0.008	0
0.4	0.378	-0.022	0.028	1	0.993	-0.007	0.007	0
1.1	0.983	-0.117	0.115	1	1.091	0.091	0.044	29

[1] Number of realizations (out of 100) with lag-1 sample correlation outside invertible range.

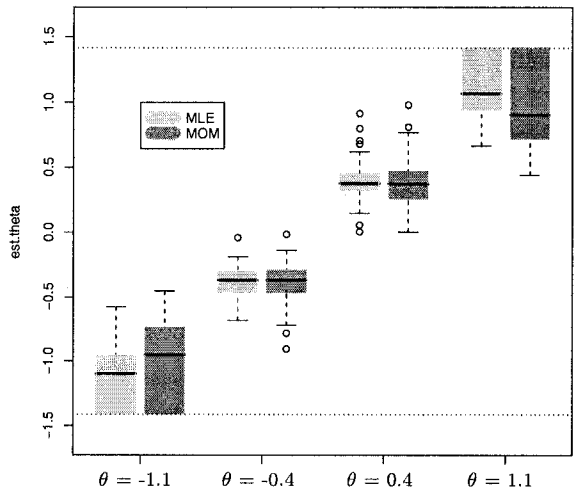
Table 4.8: Method of Moments Estimates from 100 simulated realizations of each MA(1) process on Rock Creek

θ	$E(\hat{\theta})$	$Bias(\hat{\theta})$	$MSE(\hat{\theta})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NI [1]
-1.1	-0.937	0.163	0.255	1	1.155	0.155	0.133	45
-0.4	-0.52	-0.12	0.22	1	0.91	-0.09	0.072	13
0.4	0.378	-0.022	0.165	1	0.959	-0.041	0.058	7
1.1	0.756	-0.344	0.357	1	1.225	0.225	0.14	29

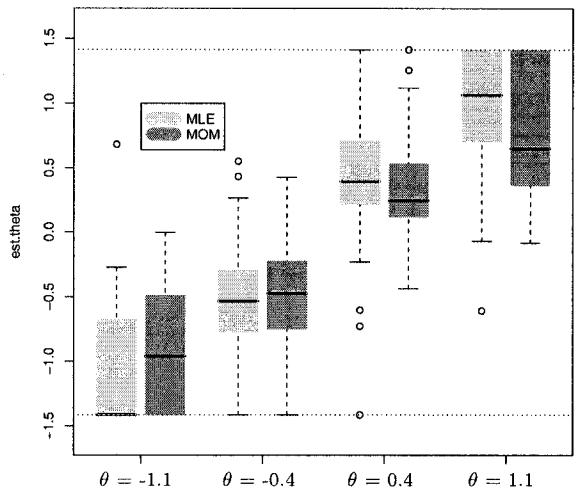
[1] Number of realizations (out of 100) with lag-1 sample correlation outside invertible range.

In Table 4.5, we see a higher mean square error for θ closer to boundaries. Upon inspection, realizations with θ close to the boundary had relatively flat likelihoods, alowing for large variation in $\hat{\theta}$. This phenomenon seems to reverse for the Rock Creek structure. With only 39 observations used for estimation, it is difficult to reason for this, other than more variation and higher biases associated with smaller sample sizes.

We notice in Tables 4.7 and 4.8 a high mean square error and biases for θ closer to boundaries. Since there are many cases where $|\hat{\rho}(1)| > 1/2\sqrt{2}$, no real valued solution for $\hat{\theta}$ can be found. Thus, $\hat{\theta}$ was defined to be the value of the closest boundary. We see this seems more likely to occur when the true θ is closer to $\pm\sqrt{2}$. In cases where $|\hat{\rho}(1)| \leq 1/2\sqrt{2}$, the corresponding $\hat{\sigma}^2$ will always be non-negative.



(a) Binary tree



(b) Rock Creek structure.

Figure 4.6: Maximum Likelihood (MLE) and Method-of-Moments (MOM) parameter estimates for 100 simulated realizations of each MA(1) process on different tree structures. Note that MOM estimates are defined to be the boundary value when the moment estimates for ρ are outside the causal region.

4.6.7 ARMA(1,1)

Realizations for different combinations of (ϕ, θ) previously mentioned with $\sigma^2 = 1$ were generated for both tree structures. Optimization was again performed using four different starting points, one from each quadrant of the region defined by the intersection of $|\phi| < \sqrt{2}$ and $|\theta| < \sqrt{2}$. Specifically, we chose $(\phi, \theta) \in \{(1, .5), (-1, .5), (1, -.5), (-1, -.5)\}$ as starting points for the optimization, and selected the resulting estimate which resulted in the highest likelihood. The resulting ACFs from averaging the sample ACF and MLE ACF over the 100 simulated realizations are shown in Figures 4.7 and 4.9.

From the plots in Figures 4.7 and 4.9, we can see that for some combinations of (ϕ, θ), the autocorrelations drop off rather quickly. This occurs when ϕ is close to $-\theta$, where the polynomials nearly cancel each other out resulting in a process that closely resembles that of white noise. This is easily seen in the plots along the back diagonal of Figure 4.7 and Figure 4.9, starting with that of $(\phi, \theta) = (-1.2, 1.1)$. As seen in the first and last column of each figure, θ generally has minimal influence on the ACF for a fixed ϕ near the boundary, with the exception being the cases when $\phi - \theta$ is closer to zero.

Parameter estimates from maximizing the Gaussian likelihood for the ARMA(1,1) models on a binary tree are shown in Table 4.9. Here we see that the maximum likelihood estimator of σ^2 tends to under-estimate the truth, albeit negligible with the larger number of reaches in the tree considered. We see cases of large biases in estimates of ϕ and θ for all cases when ϕ is close to $-\theta$, the cases in which the resulting process is close to white noise. This large bias directly impacts the mean square error, which is unusually high relative to the other ARMA(1,1) models considered.

From Figure 4.8, there tends to be a negative correlation between $\hat{\phi}$ and $\hat{\theta}$, which is more prominent along the back diagonal starting in the lower left corner. It these

Table 4.9: Maximum Likelihood Estimates from 100 simulated realizations of each ARMA(1,1) process on a Binary Tree

ϕ	$E(\hat{\phi})$	Bias($\hat{\phi}$)	MSE($\hat{\phi}$)	θ	E($\hat{\theta}$)	Bias($\hat{\theta}$)	MSE($\hat{\theta}$)	σ^2	E($\hat{\sigma^2}$)	Bias($\hat{\sigma^2}$)	MSE($\hat{\sigma^2}$)
-1.2	-1.194	0.006	0.001	-1.1	-1.205	-0.105	0.063	1	0.937	-0.063	0.027
-0.6	-0.624	-0.024	0.016	-1.1	-1.139	-0.039	0.102	1	0.955	-0.045	0.029
0.6	0.449	-0.151	0.219	-1.1	-1.013	0.087	0.3	1	0.952	-0.048	0.016
1.2	0.345	-0.855	1.824	-1.1	-0.225	0.875	1.983	1	0.933	-0.067	0.015
-1.2	-1.203	-0.003	0.002	-0.4	-0.44	-0.04	0.048	1	0.955	-0.045	0.019
-0.6	-0.606	-0.006	0.055	-0.4	-0.395	0.005	0.123	1	0.962	-0.038	0.014
0.6	0.263	-0.337	0.902	-0.4	-0.067	0.333	1.084	1	0.945	-0.055	0.014
1.2	1.21	0.01	0.011	-0.4	-0.44	-0.04	0.087	1	0.965	-0.035	0.013
-1.2	-1.192	0.008	0.013	0.4	0.413	0.013	0.096	1	0.971	-0.029	0.017
-0.6	-0.406	0.194	0.73	0.4	0.205	-0.195	0.851	1	0.947	-0.053	0.011
0.6	0.607	0.007	0.054	0.4	0.419	0.019	0.108	1	0.965	-0.035	0.011
1.2	1.19	-0.01	0.003	0.4	0.483	0.083	0.076	1	0.959	-0.041	0.022
-1.2	-0.41	0.79	1.651	1.1	0.324	-0.776	1.708	1	0.957	-0.043	0.01
-0.6	-0.469	0.131	0.203	1.1	1.063	-0.037	0.244	1	0.934	-0.066	0.019
0.6	0.624	0.024	0.016	1.1	1.132	0.032	0.092	1	0.954	-0.046	0.025
1.2	1.197	-0.003	0.002	1.1	1.129	0.029	0.075	1	0.967	-0.033	0.029

cases, we see that plots generally follow a $\hat{\phi} \approx -\hat{\theta}$, which results in estimation from a process that closely resembles white noise. This is consistent with their autocorrelation functions which indicate a very small dependence for all lags.

We also see from Figure 4.8 a generally greater variation in $\hat{\theta}$, which is believed to be a function of the small influence of θ on the likelihood resulting in a flat objective function in the θ direction. Although this neighborhood is large, likelihood plots revealed that the true parameters were generally in this neighborhood. When the true ϕ is closer to $\pm\sqrt{2}$, we see much more gradient in the likelihood, and consequently less variation in $\hat{\phi}$.

The autocorrelation functions for the ARMA(1,1) models on Rock Creek generally follow the same patterns as those on a binary tree, but with noticeable differences in the sample ACF. In this network with many fewer reaches, it is obvious that the sample ACF deviates from the truth, and is more distinguishable when $\phi > 0$. We also see that the sample ACF is often biased toward zero at lower lags, yet not at higher lags where the effect is expected to be worse.

We see from Table 4.10 that parameter estimates generally have biases larger in magnitude and corresponding mean square errors than those on the binary tree. This is believed to be a function of the sample size rather than the tree structure, although we have neither addressed nor determined any influence of tree structure on variation in parameter estimates, as it is left as an area of future research. We see the same large biases (and mean square errors) as we did on the binary tree for parameter estimates when the true ϕ is close to $-\theta$.

In Figure 4.10, again we see the same general tendencies as in Figure 4.8 for the binary tree. The greater variation in parameter estimates in Table 4.10 is clear from these plots where the clusters all appear much bigger.

Table 4.10: Maximum Likelihood Estimates from 100 simulated realizations of each ARMA(1,1) process on Rock Creek

ϕ	$E(\hat{\phi})$	Bias($\hat{\phi}$)	MSE($\hat{\phi}$)	θ	E($\hat{\theta}$)	Bias($\hat{\theta}$)	MSE($\hat{\theta}$)	σ^2	E($\hat{\sigma^2}$)	Bias($\hat{\sigma^2}$)	MSE($\hat{\sigma^2}$)
-1.2	-1.194	0.006	0.011	-1.1	-1.051	0.049	0.296	1	0.951	-0.049	0.124
-0.6	-0.734	-0.134	0.125	-1.1	-0.853	0.247	0.607	1	0.943	-0.057	0.091
0.6	-0.013	-0.613	1.006	-1.1	-0.536	0.564	1.497	1	0.879	-0.121	0.071
1.2	-0.01	-1.21	2.346	-1.1	0.176	1.276	2.942	1	0.842	-0.158	0.088
-1.2	-1.189	0.011	0.019	-0.4	-0.571	-0.171	0.445	1	0.843	-0.157	0.131
-0.6	-0.632	-0.032	0.288	-0.4	-0.441	-0.041	0.798	1	0.768	-0.232	0.123
0.6	0.247	-0.353	1.066	-0.4	-0.042	0.358	1.36	1	0.844	-0.156	0.072
1.2	1.107	-0.093	0.144	-0.4	-0.36	0.04	0.804	1	0.812	-0.188	0.107
-1.2	-1.15	0.05	0.088	0.4	0.382	-0.018	0.684	1	0.804	-0.196	0.13
-0.6	-0.136	0.464	1.085	0.4	-0.072	-0.472	1.488	1	0.834	-0.166	0.086
0.6	0.599	-0.001	0.261	0.4	0.466	0.066	0.818	1	0.851	-0.149	0.103
1.2	1.205	0.005	0.013	0.4	0.518	0.118	0.497	1	0.781	-0.219	0.121
-1.2	-0.103	1.097	2.176	1.1	0.004	-1.096	2.468	1	0.842	-0.158	0.087
-0.6	0.069	0.669	1.17	1.1	0.379	-0.721	1.788	1	0.952	-0.048	0.068
0.6	0.707	0.107	0.14	1.1	0.896	-0.204	0.625	1	0.887	-0.113	0.122
1.2	1.207	0.007	0.015	1.1	1.016	-0.084	0.345	1	0.916	-0.084	0.111

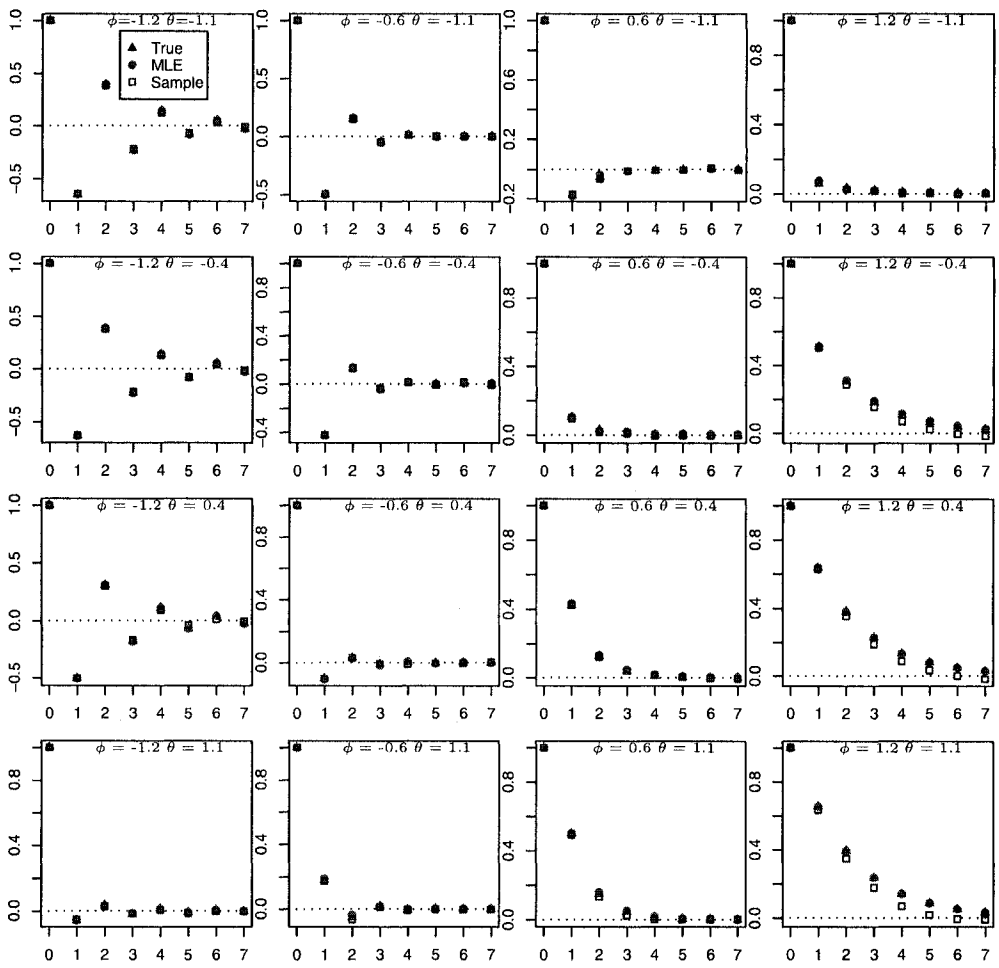


Figure 4.7: True, estimated (MLE), and empirical ACF for each ARMA(1,1) process on a binary tree. The MLE and empirical ACF displayed result from averaging the corresponding function over the 100 simulated realizations.

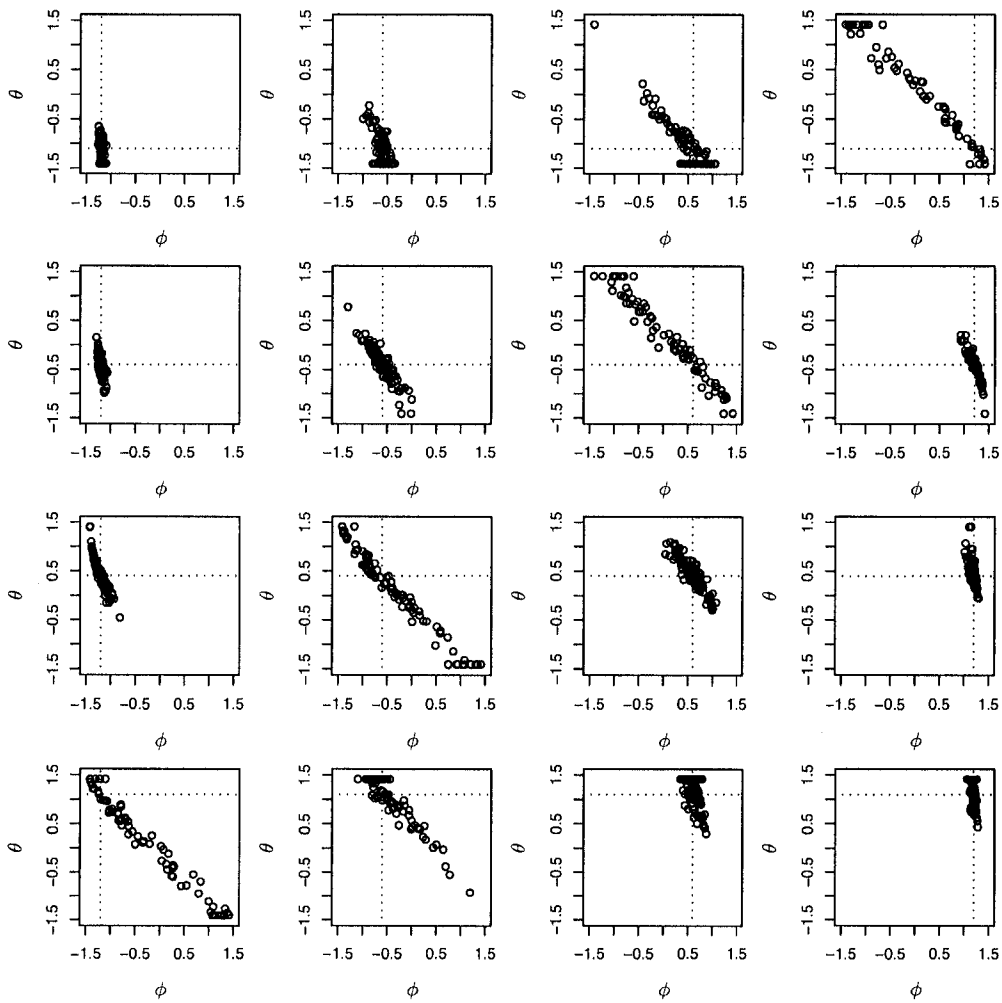


Figure 4.8: Scatter Plot of parameter estimates for 100 simulated realizations of each ARMA(1,1) process on a Binary Tree. True parameter values are indicated by the dotted lines.

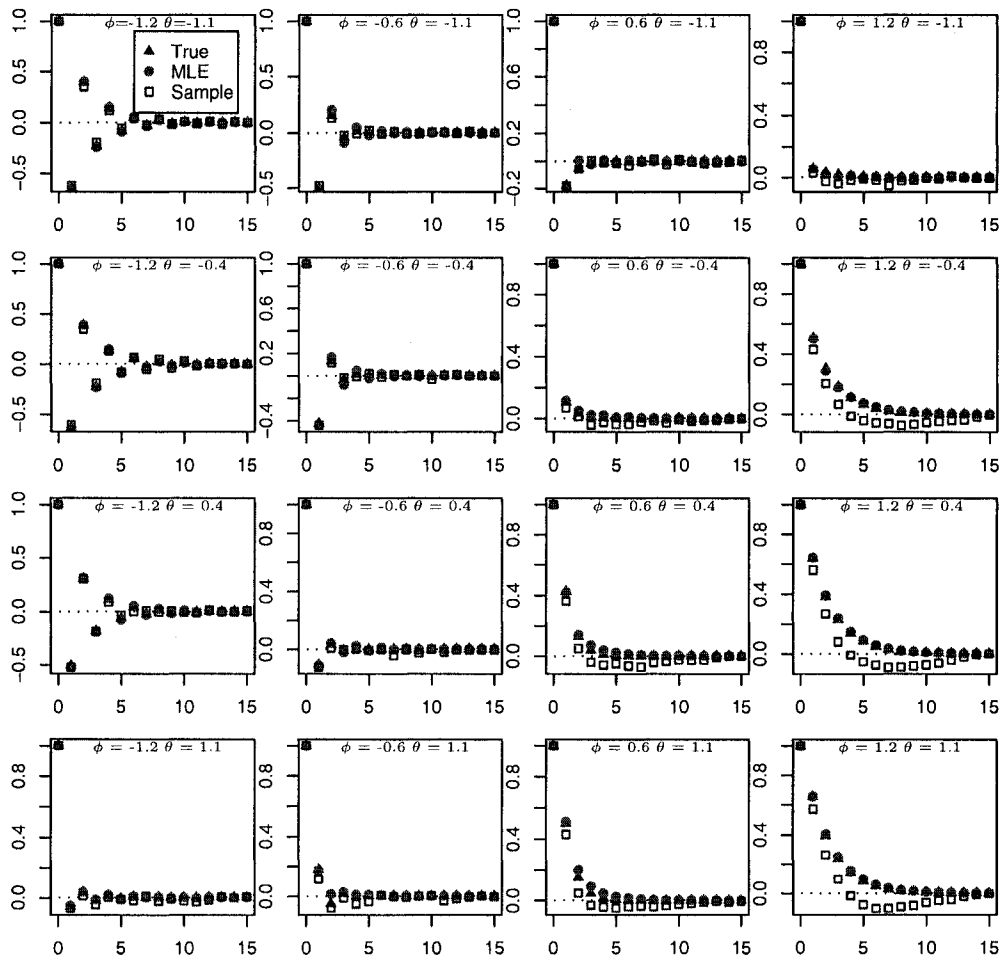


Figure 4.9: True, estimated (MLE), and empirical ACF for each ARMA(1,1) process on Rock Creek. The MLE and empirical ACF displayed result from averaging the corresponding function over the 100 simulated realizations.

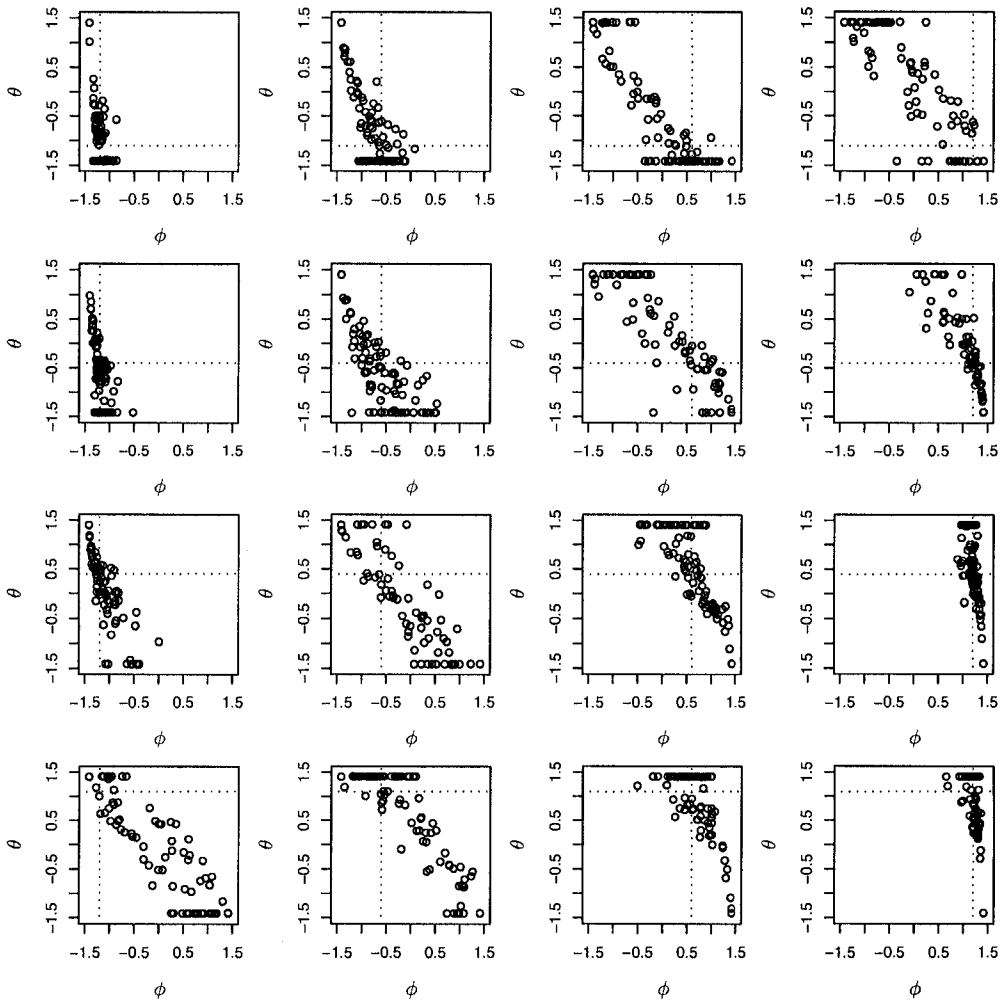


Figure 4.10: Scatter Plot of parameter estimates for 100 simulated realizations of each ARMA(1,1) process on Rock Creek. True parameter values are indicated by the dotted lines.

4.6.8 Second Order Autoregressive: AR(2)

For the AR(2) models considered, the white noise variance was again $\sigma^2 = 1$. Different pairs (ϕ_1, ϕ_2) are considered to represent different covariance structures over the stream network, namely, $(1, 0.25), (-0.75, 0.25)$, and $(0.5, -1.5)$. The first two represent causal AR(2) models with real roots, whereas the last represents a causal AR(2) with complex roots.

Yule-Walker estimates for ϕ_1 and ϕ_2 were obtained by substituting $\hat{\gamma}(h)$ from the sample ACVF into the relation

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \begin{bmatrix} \gamma(0)/2 & \gamma(1)/2 \\ \gamma(1)/2 & \gamma(0)/4 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(1) \\ \gamma(2) \end{bmatrix}$$

which results from (4.24). The estimate for σ^2 was then computed directly from (4.25).

Maximum Likelihood estimates were obtained numerically from starting points of $(\phi_1, \phi_2) \in \{(0.5, 0.7), (-0.5, 0.7), (1.0, -1.25), (-1.0, -1.25)\}$. Additionally, if the Yule-Walker estimates fell in the causal region defined in Figure 4.1, they too were used as an initial starting point.

In the previous models, we used appropriate transformations to allow for an unconstrained optimization that guaranteed estimates from a causal/invertible model. We have not found any such transformations for the AR(2) model. To achieve causality, the objective function was set to an arbitrarily large number if the iterative estimates fell outside the causal boundary. This rule kept the routine within the causal region. As seen in Tables 4.13 and 4.14, there are cases where Yule-Walker estimates are not guaranteed to be inside the causal region. Results from 100 realizations under the tree models considered are tabled below.

The behavior of the ACF under different models is shown in Figure 4.11, where the sample ACF and MLE ACF are averages over the corresponding ACF for each of the 100 simulated realizations. The ACF decays with increasing lag. When the

autoregressive polynomial $\phi(z/2)$ has complex roots, the ACF resembles a dampening sinusoid. When the modulus of the roots associated with $\phi(z/2)$ are close to $\sqrt{2}$, the dampening is slower. This differs from that in time series where the dampening is slower when r is close to 1. Here, we require that r be at least $\sqrt{2}$ for the AR(2) to be causal.

For the estimates obtained via maximum likelihood, Table 4.11 shows much smaller biases and mean-square errors associated with the binary tree than from Rock Creek, again attributed to the much larger sample size. Consequently, there are many more transitions at each lag h that are available for parameter estimation. We also notice that there tends to be a negative bias associated with all estimates from the Rock Creek structure.

The variation within and correlation between $\hat{\phi}_1$ and $\hat{\phi}_2$ from maximum likelihood can be seen in Figures 4.12 and 4.13. In comparing the maximum likelihood estimators with those from the Yule-Walker equations, we see much less variation in those obtained via maximizing the Gaussian likelihood. We also see a more obvious correlation between $\hat{\phi}_1$ and $\hat{\phi}_2$ for the models with (true) real roots, correlation that is more prominent in the estimates from maximum likelihood.

We also note that even though the optimization algorithm was restricted to the causal region defined in Figure 4.1, the optimization converged for every realization within each model. Without this restriction, the optimization periodically jumped outside of the causal boundaries, a likely consequence of relatively flat likelihoods for the models selected.

Although not shown in the plot, we periodically observed unusual YW estimates such as $\hat{\phi}_1 \approx -35$ with corresponding $\hat{\phi}_2 \approx 51$ when the true values were $\phi_1 = 1.0$ and $\phi_2 = 0.25$. These large anomalies contribute to the large biases and mean square error. Although there are many more occurrences of these anomalies in the model with complex roots where $\phi_1 = 0.5$ and $\phi_2 = -1.5$, these estimates tended to be

Table 4.11: Maximum Likelihood Estimates from 100 simulated realizations of each AR(2) process on a Binary Tree

ϕ_1	$E(\hat{\phi_1})$	$Bias(\hat{\phi_1})$	$MSE(\hat{\phi_1})$	ϕ_2	$E(\hat{\phi_2})$	$Bias(\hat{\phi_2})$	$MSE(\hat{\phi_2})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
1	0.993	-0.007	0.017	0.25	0.255	0.005	0.05	1	0.985	-0.015	0.017	0
-0.75	-0.748	0.002	0.014	0.25	0.264	0.014	0.05	1	0.967	-0.033	0.012	0
0.5	0.512	0.012	0.01	-1.5	-1.504	-0.004	0.009	1	0.969	-0.031	0.023	0

[1] Number of realizations (out of 100) with non-convergence in Maximum Likelihood estimation.

Table 4.12: Maximum Likelihood Estimates from 100 simulated realizations of each AR(2) process on Rock Creek

ϕ_1	$E(\hat{\phi_1})$	$Bias(\hat{\phi_1})$	$MSE(\hat{\phi_1})$	ϕ_2	$E(\hat{\phi_2})$	$Bias(\hat{\phi_2})$	$MSE(\hat{\phi_2})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
1	0.985	-0.015	0.1	0.25	0.216	-0.034	0.311	1	0.869	-0.131	0.126	0
-0.75	-0.767	-0.017	0.128	0.25	0.202	-0.048	0.431	1	0.781	-0.219	0.13	0
0.5	0.48	-0.02	0.144	-1.5	-1.546	-0.046	0.249	1	0.71	-0.29	0.405	0

[1] Number of realizations (out of 100) with non-convergence in Maximum Likelihood estimation.

Table 4.13: Yule-Walker Estimates from 100 simulated realizations of each AR(2) process on a Binary Tree

ϕ_1	$E(\hat{\phi_1})$	$Bias(\hat{\phi_1})$	$MSE(\hat{\phi_1})$	ϕ_2	$E(\hat{\phi_2})$	$Bias(\hat{\phi_2})$	$MSE(\hat{\phi_2})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
1	0.894	-0.106	0.411	0.25	0.287	0.037	0.97	1	1.076	0.076	0.27	1
-0.75	-0.712	0.038	0.026	0.25	0.256	0.006	0.087	1	0.99	-0.01	0.022	0
0.5	0.603	0.103	0.038	-1.5	-1.533	-0.033	0.089	1	0.868	-0.132	0.277	6

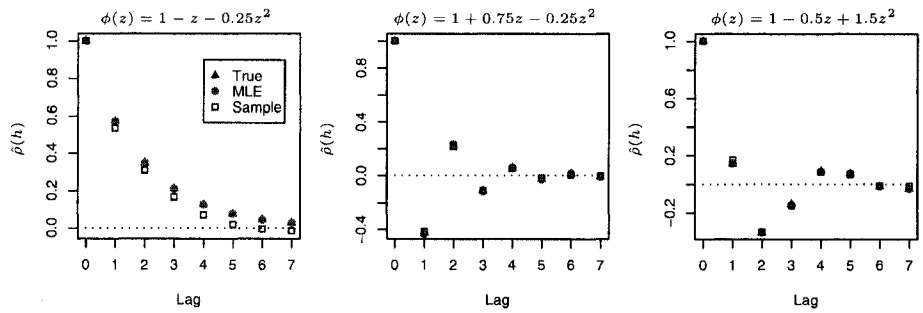
[1] Number of realizations (out of 100) with non-causal YW estimates.

Table 4.14: Yule-Walker Estimates from 100 simulated realizations of each AR(2) process on Rock Creek

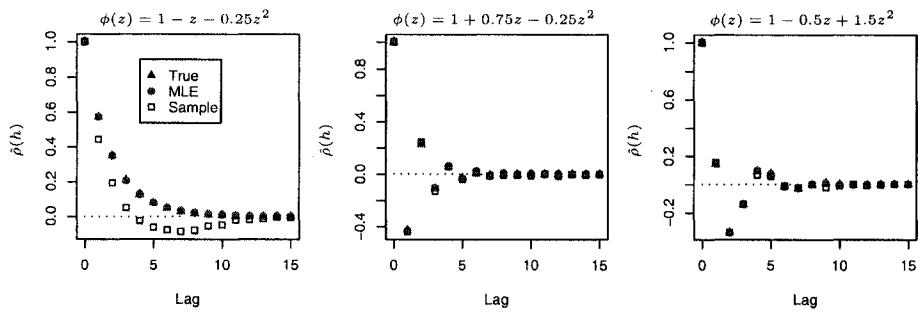
ϕ_1	$E(\hat{\phi_1})$	$Bias(\hat{\phi_1})$	$MSE(\hat{\phi_1})$	ϕ_2	$E(\hat{\phi_2})$	$Bias(\hat{\phi_2})$	$MSE(\hat{\phi_2})$	σ^2	$E(\hat{\sigma^2})$	$Bias(\hat{\sigma^2})$	$MSE(\hat{\sigma^2})$	NC [1]
1	0.657	-0.343	13.782	0.25	0.232	-0.018	27.774	1	1.645	0.645	19.18	6
-0.75	-0.955	-0.205	6.413	0.25	-0.1	-0.35	12.98	1	0.501	-0.499	3.348	9
0.5	0.548	0.048	0.208	-1.5	-1.567	-0.067	0.426	1	0.566	-0.434	1.621	25

[1] Number of realizations (out of 100) with non-causal YW estimates.

fairly close to the causal boundary, whereas those for the other models tended to be much more extreme, thus having a greater impact on the estimates of bias and mean square error.

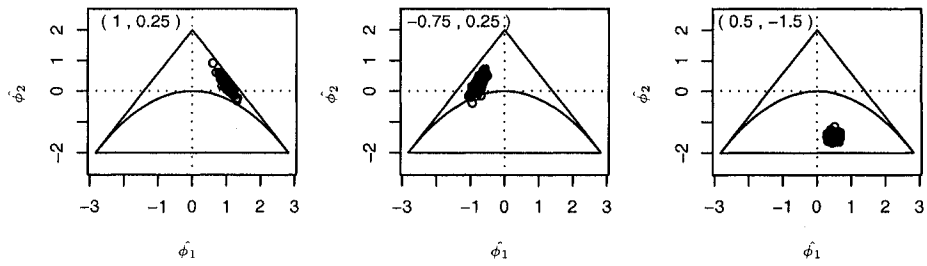


(a) Binary Tree

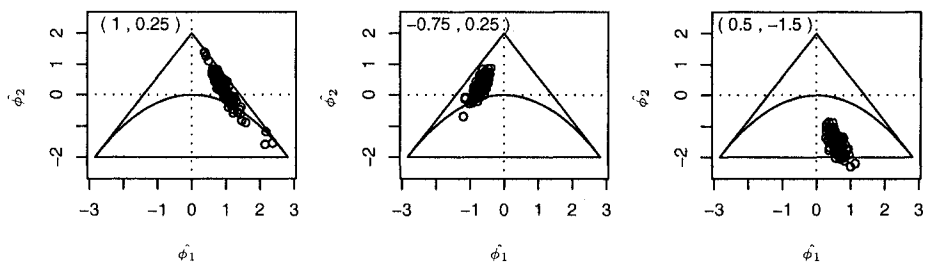


(b) Rock Creek structure.

Figure 4.11: True, estimated (MLE), and empirical ACF for each AR(2) process on different tree structures. The MLE and empirical ACF displayed result from averaging the corresponding function over the 100 simulated realizations.

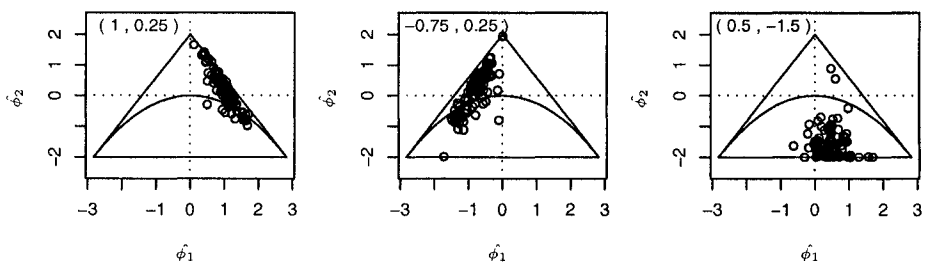


(a) Maximum Likelihood

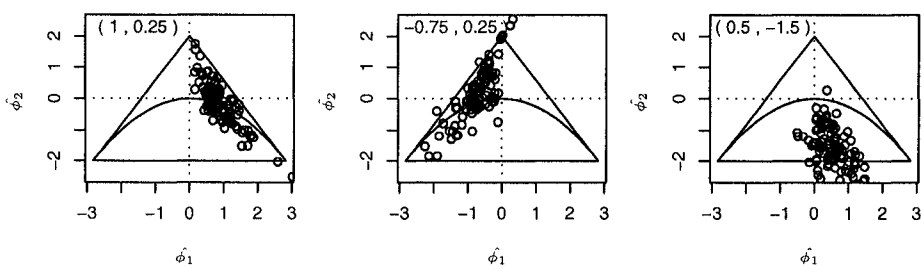


(b) Yule-Walker

Figure 4.12: Scatter plots of $(\hat{\phi}_1, \hat{\phi}_2)$ for 100 simulated realizations of each AR(2) process on a binary tree.



(a) Maximum Likelihood



(b) Yule-Walker

Figure 4.13: Scatter plots of $(\hat{\phi}_1, \hat{\phi}_2)$ for 100 simulated realizations of each AR(2) process on Rock Creek.

4.6.9 Model Fit to Rock Creek Data

Data for numerous study variables are available for river networks in Montgomery County, Maryland, from the state's Department of Environmental Protection. This agency oversees biological monitoring stations throughout the county and in some instances other counties where streams enter into Montgomery County. In general, the stations are randomly located within each of the watersheds. Without the use of Geographic Information Software, often referred to as GIS, there is no good estimate of how much distance water must travel between monitoring stations. We ignore this specific information for now, although future research may involve the use of such information to appropriately weight specific observations based on some sort of distance measure.

The data of interest here are fish habitat information at each of the stations. Variables such as instream cover, channel attributes, and suspended sediment are available for monitoring stations throughout the county at a variety of time points. For simplicity, we selected only a small drainage to work with, that of Upper and Lower Rock Creek. Many reaches have multiple monitoring stations that may have collected data during summer and winter months since the mid 1990s. The geometry of this specific watershed and distribution of observation stations can be seen in Figure 4.14. There are 39 reaches total, however, only 35 have at least 1 monitoring station.

We consider instream cover score in this example, which is a measure of the percent of covered areas under water in which fish can hide, and ranges from 0 to 20. Although these are discrete scores, we treat these values as if they were continuous for this analysis. Multiple observations per reach are considered conditionally independent given the state of the reach even though they may be from different monitoring stations or different time points. We fit four models to this data: AR(1),

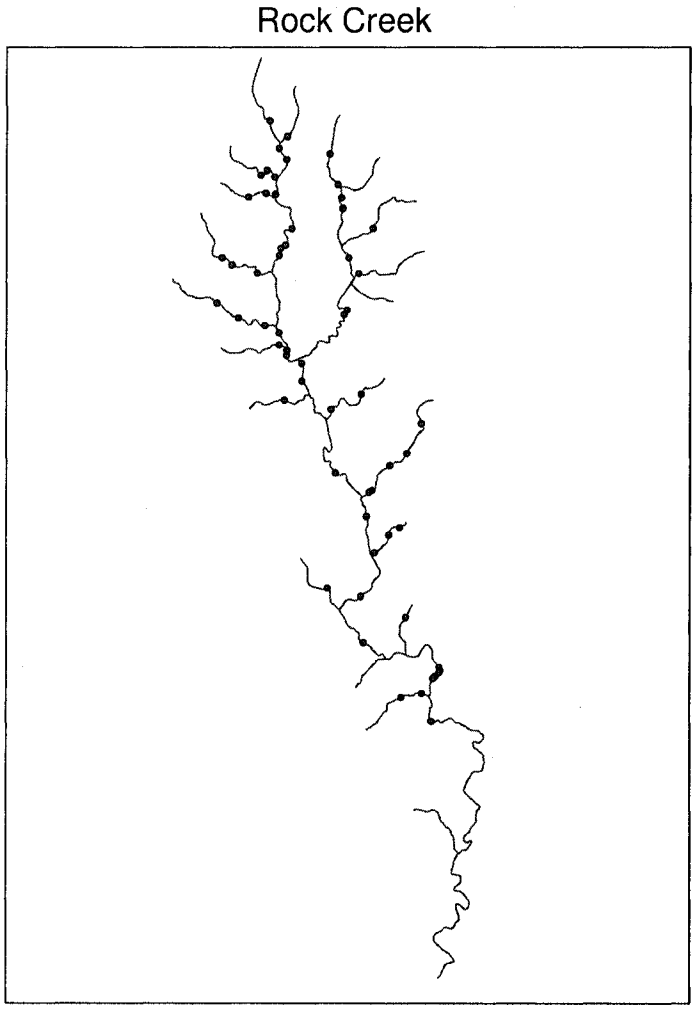


Figure 4.14: Geometric structure and distribution of monitoring stations for Upper and Lower Rock Creek.

MA(1), ARMA(1,1), and AR(2), each with added noise to account for the multiple observations per reach. Thus, the observation equation for the ARMA models is redefined such that $\mathbf{W}(k)$ is no longer deterministically zero. The observation equation is now defined by (2.1) with

$$\mathbf{W}(k) \sim N\left(\mathbf{0}, \sigma_w^2 I_{n_k \times n_k}\right).$$

where n_k is the number of observations for reach k.

We obtain a sample ACF through a sampling approach. Here we randomly select one observation per reach to obtain a single univariate realization over the network, and calculate a sample ACF for this realization. We repeat this process 100 times, with the average at each lag then taken to be the sample ACF at that lag for the process on the network.

Using a variation of standard permutation tests (Mielke and Berry, 2001), we empirically compute bounds at each lag as if the data were white noise. To do so, we permute the data for the entire network to assign new values to each reach eliminating any structure to the data. A sample ACF via the sampling approach previously described was then calculated. This process was repeated 250 times resulting in a sample of possible autocorrelations at each lag for unstructured data. We define the bounds for a white noise process to be the 2.5^{th} and 97.5^{th} percentiles at each lag.

We see from Figure 4.15 that there seems to be a small amount positive autocorrelation at the first two lags, which is believed to under estimate the truth because of the downward bias seen in the sample ACFs in the simulation study, particularly on the Rock Creek structure. We fit four models to this data via Maximum Likelihood, and generate an estimated autocorrelation function based on the model fit. For assurance that the estimates truely maximize the likelihood, the function was first evaluated at a grid of possible parameter values. We then chose the twenty grid

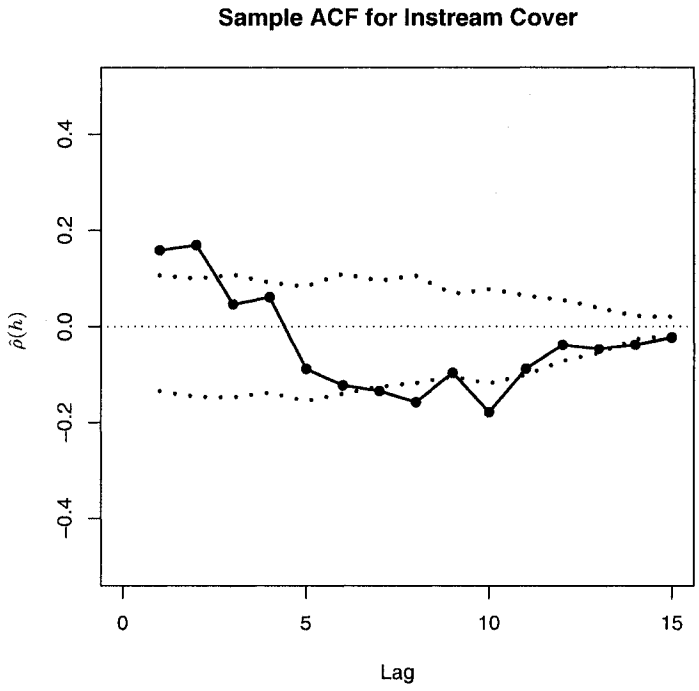


Figure 4.15: Sample Autocorrelation Function for mean corrected instream cover values on Rock Creek with empirical white noise bounds.

points that resulted in the highest likelihood to be used as starting values for the optimization. From those optimizations, the estimates were those with the largest likelihood such that the optimization converged.

The autocorrelation function of each fitted model was adjusted accordingly due to the measurement variation. In doing so, we have an estimated autocorrelation which is representative of the observations rather than a function strictly for the underlying process. These fitted autocorrelation functions are seen in Figure 4.16. In Figure 4.16, we see similar discrepancies between the MLE and empirical ACF to those identified in the simulation studies.

Another diagnostic utilized for model selection resulted from a form of a parametric bootstrap (Lange, 1998). One hundred realizations under the fitted model

Table 4.15: MA(1) fit to instream cover on Rock Creek

Parameter	Estimate	Bootstrap Standard Error	Bootstrap Interval
σ_z^2	59.61	21.32	(40.59, 116.99)
σ_w^2	7.86	0.92	(6.35, 9.86)
θ	1.41	0.34	(0.44, 1.41)

were generated over the stream network, and a sample ACF for each was determined. From this, we determine if the observed sample ACF is representative of a sample ACF for data generated by the fitted model. We see a set of possible sample ACFs for each model fit in Figure 4.16. Of the models considered, the MLE ACF is more consistent with sample ACFs of an MA(1) process than with a process with non-zero autoregressive parameters.

We obtain estimated standard errors and confidence interval approximations via the same parametric bootstrap in the above diagnostic. For each realization generated by the parametric model fit, new parameter estimates were obtained. The bootstrap standard error reported in Table 4.15 is the standard deviation of the parameter estimates obtained from the 100 bootstrap realizations. The interval estimate is obtained from the 2.5^{th} and 97.5^{th} percentiles providing bounds for the middle 95% of the bootstrap estimates. We see from Table 4.15 that our estimate of θ is not far from $\sqrt{2}$, and seems significantly different from zero. This may be an indication of some trend that has yet to be removed, as is sometimes the case in time series. We do find in exploration of the data a decreasing trend in instream cover with progression downstream.

An MA(2) was fit to the original instream cover values. Although the properties of a formal likelihood ratio test have yet to be determined, the ratio of likelihoods indicates minimal gain from modeling the more complicated MA(2). Rather than focusing on a more complicated model, we attempted to remove the spatial dependence by regressing instream cover on landscape and other appropriate covariates.

Table 4.16: MA(1) fit to residuals on Rock Creek

Parameter	Estimate	Bootstrap Standard Error	Bootstrap Interval
σ_z^2	0.53	0.47	(0.05, 2.07)
σ_w^2	7.86	0.89	(6.35, 9.64)
θ	-1.08	1.05	(-1.41, 1.41)

Available information such as month, year, and elevation are specific to monitoring station, while land use and gradient associated with stream bed are reach specific. Other information such as percent of local basin that is forested or cultivated was also available.

A linear model was fit using these covariates, where we then looked for autocorrelation in the residuals from this regression model. We see the residuals seem to be representative of white noise from Figure 4.17.

We fit the MA(1) the these residuals, and obtained an interval estimate as well as standard errors via the parametric bootstrap. Results are found in Table 4.16. The estimate of θ indicates negative dependence, yet our bootstrap interval indicates that this estimate is not significantly different from zero. We also see this in Figure 4.18.

Thus, we see that the residuals do resemble a white noise process, and that the autocorrelation in instream cover was removed given appropriate landscape characteristics.

By example, we have shown the ability to fit several simple ARMA models to actual data collected on Rock Creek. We have obtained estimates of model parameters and autocorrelation functions. Although we have not been able to address the adequecy of the MA(1) model fit to instream cover, we have shown that the autocorrelation can possibly be removed given the appropriate landscape covariates. Tools to assess model adequecy and model selection are left as an open area of research.

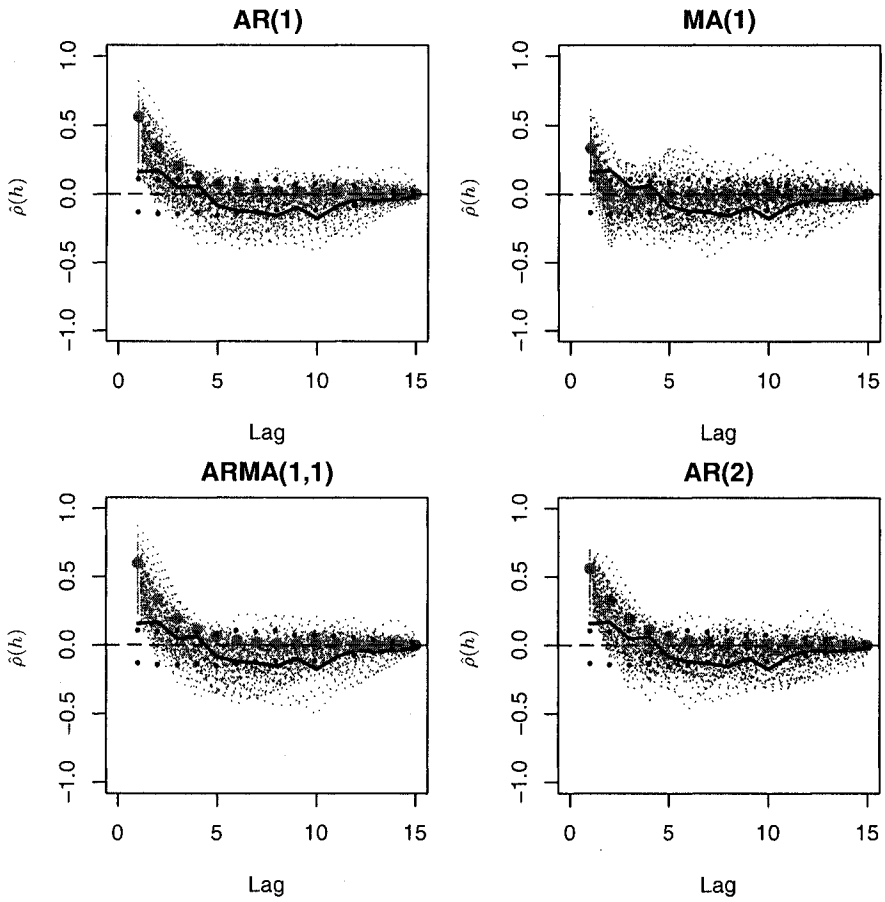


Figure 4.16: Fitted Autocorrelation Functions (MLE) for mean corrected instream cover values on Rock Creek identified by red circles. Grey lines indicate sample ACFs of parametric bootstrap realizations whereas the solid black line is the sample ACF of the observed data. Empirical white noise bounds are indicated by black small black circles.

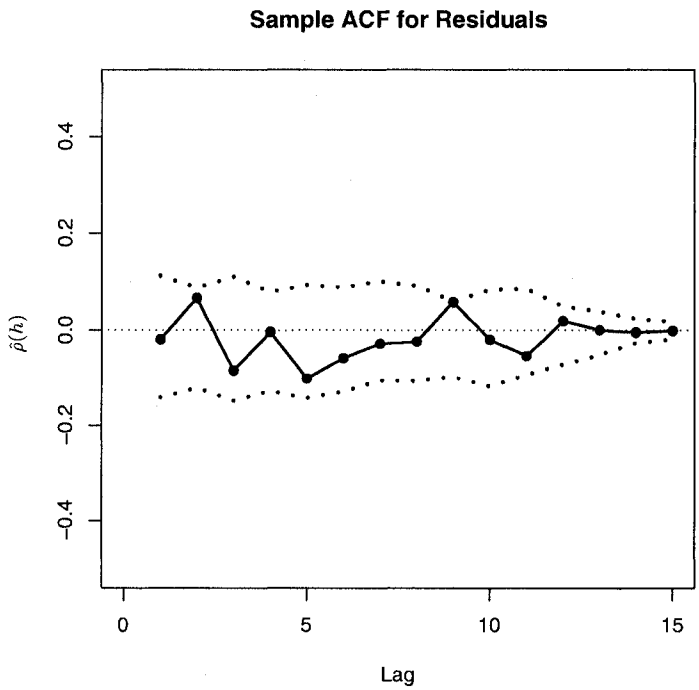


Figure 4.17: Sample Autocorrelation Function for the residuals from a linear model fit to instream cover on Rock Creek with empirical white noise bounds.

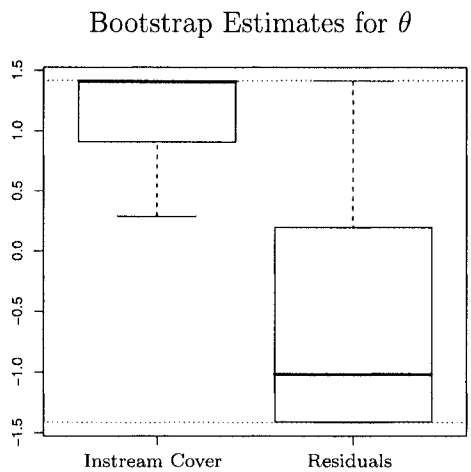


Figure 4.18: Parametric Bootstrap results for an MA(1).

Chapter 5

NON-STATIONARY MODELS

In contrast with the ARMA models introduced in the previous chapter, models such that first and second moments can depend on reach location are said to be nonstationary. Such models are easily formulated when model parameters depend on location. Models can also be defined explicitly in terms of components of interest such as trend or seasonality, which although they are unobservable, have direct interpretation. These models are referred to as *structural* models in a time series context (see Harvey, 1989, p.44). The components of these models are not seen as deterministic, rather each is driven by random disturbances resulting in a nonstationary model. The stochastic setting allows for a series to respond to general changes in behavior, especially when explanatory information can not be directly measured.

From the class of possible structural time series models, we consider only stochastic trend models on a stream network, although addition of other stochastic components is feasible. We develop a network analogue of Random Walk with noise and Local Linear Trend models. We define a discrete smoothing spline on a stream network using a special case of these stochastic trend models, and apply it to data on Rock Creek where the initial conditions are estimated via a concentrated likelihood. The smoothing spline is then obtained via the Kalman recursions developed in Chapter 2.

Although more general nonstationary models can allow for much flexibility, parameter estimates for even simple models can be difficult to obtain and may be

unreliable as was seen in §3.5. In this example we demonstrated that estimation even with a relatively simple model can be problematic when the network is small such that it has few transitions from reach to reach and missing data are present. We do not pursue these nonstationary models any further in this dissertation.

We begin this chapter with the introduction of stochastic trend models with an application on a stream network.

5.1 Stochastic Trend Models

The stochastic trend models considered are the network analogues of a Random Walk plus Noise (RW+N) and Local Linear Trend (LLT). In these cases, the variances increase with progression downstream. In either case, a difference operation can be performed to obtain a stationary series over a network. However, because of the branching structure, the number of observations in the resulting tree structure may be dramatically reduced by differencing. Differencing is further complicated by missing values. We therefore avoid differencing, and instead use the Kalman Filter to construct a likelihood to be used as a criterion function for purposes of estimation and prediction.

We initially consider the RW+N, and later show it as a special case of the LLT. The Exact Kalman recursions developed in §2.5.2 are applied to the RW+N, as the tree structures considered will allow for this simple model.

5.1.1 Random Walk plus Noise

The Random Walk plus Noise (RW+N), or local level model, over the stream network is defined by

$$Y(k) = X(k) + W(k) \tag{5.1}$$

$$X(k) = \frac{1}{2}\left(X(u_1) + X(u_2)\right) + V(k) \tag{5.2}$$

where $W(k) \sim N(0, \sigma_w^2)$ and $V(k) \sim N(0, \sigma_v^2)$ are uncorrelated white noise. We further assume that $X(k)$, for each first order k, is a random draw from $N(0, \tau^2)$. Then using the recursive relationship in $X(k)$, we see that

$$\begin{aligned}
\text{Var}(X(k)) &= \frac{1}{4}\text{Var}(X(u_1)) + \frac{1}{4}\text{Var}(X(u_2)) + \text{Var}(V(k)) \\
&= \frac{1}{4}\left\{\frac{1}{4}\text{Var}(X_{21}(k)) + \frac{1}{4}\text{Var}(X_{22}(k) + \text{Var}(V_{11}(k))\right\} \\
&\quad + \frac{1}{4}\left\{\frac{1}{4}\text{Var}(X_{23}(k)) + \frac{1}{4}\text{Var}(X_{24}(k) + \text{Var}(V_{12}(k))\right\} \\
&\quad + \text{Var}(V(k)) \\
&\quad \vdots \\
&= \sigma_v^2 \sum_{k' \in \mathcal{H}_{(k]}} \left(\frac{1}{4}\right)^{|k-k'|} + \tau^2 \sum_{k' \in \mathcal{F}_k} \left(\frac{1}{4}\right)^{|k-k'|} \qquad (5.3)
\end{aligned}$$

where $\mathcal{H}_{(k]}$ denotes the set of all higher order reaches that are upstream of k, including reach k, and $\mathcal{F}_k$ denotes the set of first order reaches that are upstream of k. From this we see that the variation depends on the number of traces back upstream over higher order reaches as well as the number of upstream first order reaches. The covariance is seen to be

$$\text{Cov}(X(k), X(k')) = \left(\frac{1}{2}\right)^{|k-k'|} \text{Var}(X(k')) \qquad (5.4)$$

for k' upstream of k. The covariance is intuitive since the path from $X(k')$ to $X(k)$ contains $X(k')$ and independent error terms. Here we see that if u_1 has more upstream generations than u_2, then $X(k)$ will be more correlated with $X(u_1)$. Furthermore, if $|k - k'|$ is large, then k' is an ancestor from many generations prior to k, and the correlation decreases.

Under the model assumed for first order reaches, the correlation in $X(k)$ is primarily based on number of generations between k and k', where the influence of σ_v^2 and τ^2 depends on the geometric structure.

Remark 5.1.1 If we modeled $X(k)$ using sums over parents rather than averages, it is easy to show that

$$\text{Var}(X(k)) = n_1\sigma_v^2 + n_2\tau^2 \tag{5.5}$$

where n_1 is the size of $\mathcal{H}_{(k]}$ and n_2 is the size of $\mathcal{F}_k$. Furthermore,

$$\text{Cov}(X(k), X(k')) = \text{Var}(X(k')) \tag{5.6}$$

for k' upstream of k. Clearly, the variance rapidly increases, and the correlation is a function of the model variances as well as the number of upstream reaches. The difference operator is easily adapted to accommodate such models.

Remark 5.1.2 A random walk defined by (5.2) is nonstationary by assumptions regarding first order states. If the variance of a first order state $X(k)$ is defined to be $2\sigma_v^2$, then $X(k)$ in (5.2) is a stationary AR(1) process with $\phi = 1$.

5.1.1.1 First Differences

First differences of $Y(k)$ in this local level model result in

$$\begin{aligned}
\nabla Y(k) &= (1-B)Y(k) \\
&= Y(k) - \frac{1}{2}\left(Y_{11}(k) + Y_{12}(k)\right) \\
&= W(k) - \frac{1}{2}\left(W_{11}(k) + W_{12}(k)\right) + X(k) - \frac{1}{2}\left(X_{11}(k) + X_{12}(k)\right) \\
&= W(k) - \frac{1}{2}\left(W_{11}(k) + W_{12}(k)\right) + V(k)
\end{aligned}$$

which is a 1-correlated process with covariance defined as

$$\gamma(h) = \begin{cases} \sigma_v^2 + \frac{3}{2}\sigma_w^2 & h = 0, \\ \frac{-1}{2}\sigma_w^2 & h = 1, \\ 0 & h \geq 2. \end{cases}$$

Using Proposition 4.5.1, $\nabla Y(k)$ is an MA(1) process with parameters found by first solving a quadratic in θ defined by

$$\frac{\theta/2}{1+\theta^2/2} = \frac{-\sigma_w^2/2}{\sigma_v^2 + (3/2)\sigma_w^2}.$$

This quadratic has two solutions, θ_1 and θ_2, where it is easily shown that $\theta_1\theta_2 = 2$. However, the solution

$$\theta_1 = \frac{-(2\sigma_v^2 + 3\sigma_w^2) + \sqrt{4\sigma_v^4 + 12\sigma_v^2\sigma_w^2 + \sigma_w^4}}{2\sigma_w^2}$$

leads to an invertible MA(1). To see this, we find that

$$\begin{aligned} |\theta_2| &= \left| \frac{(-1)\left((2\sigma_v^2 + 3\sigma_w^2) + \sqrt{4\sigma_v^4 + 12\sigma_v^2\sigma_w^2 + \sigma_w^4}\right)}{2\sigma_w^2} \right| \\ &\geq \frac{3\sigma_w^2}{2\sigma_w^2} \\ &> \sqrt{2} \end{aligned}$$

and since $\theta_1\theta_2 = 2$, it must be that

$$\begin{aligned} |\theta_1| &= \frac{2}{|\theta_2|} \\ &< \sqrt{2}. \end{aligned}$$

5.1.1.2 Exact Kalman Filter and Smoother

Assuming no missing data, the exact Kalman recursions can be easily applied to any tree structure with no missing data under the Random Walk plus Noise model. Each basin is defined by a singe first order reach. Although the initial prediction variance is diffuse, as long as data are observed, the filtering variance is zero. Hence, the prediction variances associated with all downstream recursions are finite.

For any first order reach, define $\Omega_k^p = \kappa$ and $\Delta_k = \kappa + \sigma_w^2$ so that

$$\begin{aligned} \Omega_{\infty,k}^p &= 1 & \Delta_{\infty,k} &= 1 \\ \Omega_{\star,k}^p &= 0 & \Delta_{\star,k} &= \sigma_w^2 \end{aligned}$$

which can be used with

$$\begin{aligned} X^p(k)^{(0)} &= 0 & \mathrm{v}^p(k)^{(0)} &= Y(k) \\ X^p(k)^{(1)} &= 0 & \mathrm{v}^p(k)^{(1)} &= 0 \end{aligned}$$

to begin using the recursions defined in §2.5.2. The first filtering step yields

$$X^f(k)^{(0)} = Y(k) \qquad \Omega_{\infty,k}^f = 0 \qquad \Omega_{\star,k}^f = \sigma_w^2$$

when $Y(k)$ is observed. Since we start the recursions furthest upstream where two first order reaches merge, using (2.34), (2.35), and (2.35), we see

$$X^p(k)^{(0)} = \frac{Y(u_1)+Y(u_2)}{2} \qquad \Omega^p_{\infty,k} = 0 \qquad \Omega^p_{\star,k} = \frac{\sigma_w^2}{2} + \sigma_v^2$$

for the first downstream prediction. If we have observations on every reach, we see that the prediction variance of any higher order reach is always finite. The usual Kalman recursions can then be applied after all first order reaches are filtered, starting at the second order reach furthest upstream.

The smoother is expressed using the alternative backwards recursions defined in §2.4.3, where care must be taken in choosing the appropriate backwards recursion in $r(u_i)$. To smooth a first order reach u_i, we note that

$$r^{(0)}(k) = r(k) \qquad N_k^{(0)} = N_k$$

from the usual formulation. Furthermore, we see that

$$L^{(0)}_{k,u_i} = 0 \qquad L^{(1)}_{k,u_i} = \frac{\sigma_w^2}{2}$$

from which we find

$$r^{(0)}(u_i) = 0 \qquad r^{(1)}(u_i) = y(u_i) + \frac{\sigma_w^2}{2} r(k).$$

By (2.36), we have the obtain the smoothed value for the first order u_i

$$X^s(u_i) = y(u_i) + \frac{\sigma_w^2}{2} r(k).$$

For the smoothed variance, we have

$$N^{(0)}_{u_i} = 0 \qquad N^{(1)}_{u_i} = 1 \qquad \frac{\sigma_w^2}{2} N(k) - \sigma_w^2$$

which can be substituted into (2.37) to obtain

$$\Omega^s_{u_i} = \sigma_w^2 - \left(\frac{\sigma_w^2}{2}\right)^2 N(k).$$

When smoothing upstream, it is possible that $r(u_i)$ for one parent may be directly calculated from the usual expression $\mathbf{r}(u_i) = G_{u_i}^T \Delta_{u_i}^{-1} \mathbf{v}(u_i) + L_{k,u_i}^T \mathbf{r}(k)$ since all terms may be independent of κ, whereas smoothing the other parent may require the modified recursion where $r^{(0)}(u_i)$ and $r^{(1)}(u_i)$ are required to compute $X^s(u_i)$. This is predominantly the case in Rock Creek since so many first order reaches merge with one of higher order.

5.1.2 Local Linear Trend

A process similar to the Random Walk plus Noise but with more correlation structure can be defined by adding a local slope component to (5.2) that is itself a random walk. The consequence of adding this term is stronger correlation between successive measurements. In the LLT models considered, independence within and between all error terms is assumed. The univariate case of a Local Linear Trend model on a stream network is defined by

$$\begin{aligned} Y(k) &= X(k) + W(k) \\ X(k) &= \frac{1}{2}\left(X(u_1) + X(u_2)\right) + M(k) + V(k) \\ M(k) &= \frac{1}{2}\left(M(u_1) + M(u_2)\right) + U(t) \end{aligned}$$

for $W(k) \sim N(0, \sigma_w^2)$, $V(k) \sim N(0, \sigma_v^2)$, and $U(k) \sim N(0, \sigma_u^2)$. We also assume that all three error terms are mutually independent for all k, and that for any first order k,

$$\begin{bmatrix} X(k) \\ M(k) \end{bmatrix} \sim N\left(\begin{bmatrix} x_0 \\ m_0 \end{bmatrix}, \begin{bmatrix} \tau_x^2 & 0 \\ 0 & \tau_m^2 \end{bmatrix} \right)$$

which implies that $\mathrm{Cov}(X(j), M(k)) = 0$ for all k and any first order j. For this model, we can define the state vector to be $\mathbf{X}(k) = [X(k), M(k)]^T$ which leads to the state-space representation

$$Y(k) = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{X}(k) + W(k)$$

$$\mathbf{X}(k) = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1/2 \end{bmatrix} \mathbf{X}(u_1) + \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1/2 \end{bmatrix} \mathbf{X}(u_2) + \begin{bmatrix} U(k)+V(k) \\ U(k) \end{bmatrix}$$

where $\{R(k)\} = \{\sigma_w^2\}$ and $Q(k) = \begin{bmatrix} \sigma_u^2+\sigma_v^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 \end{bmatrix}$ for all higher order k. The variances and covariances in this non-stationary model rapidly accumulate with progression through the network as more error terms are added with each confluence of reaches.

Some preliminary definitions and derivations can aid in formulating a variance for $X(k)$. Since $M(k)$ is simply a random walk, (5.3) and (5.4) apply directly to $M(k)$. Considering second moments, we see that

$$\begin{aligned}
\mathrm{Cov}(X(k),M(k)) &= \mathrm{Cov}(\frac{1}{2}X(u_1)+\frac{1}{2}X(u_2)+M(k)+V(k),M(k)) \\
&= \frac{1}{2}\mathrm{Cov}(X(u_1),M(k))+\frac{1}{2}\mathrm{Cov}(X(u_2),M(k)) \\
&\quad +\mathrm{Var}(M(k))+\mathrm{Cov}(V(k),M(k)) \\
&\quad \vdots \\
&= \sum_{k'\in\mathcal{H}_{(k)}} \left(\frac{1}{2}\right)^{|k-k'|} \mathrm{Cov}(M(k'),M(k)) \\
&= \sum_{k'\in\mathcal{H}_{(k)}} \left(\frac{1}{4}\right)^{|k-k'|} \mathrm{Var}(M(k')) \\
&= \sigma_u^2 \sum_{k'\in\mathcal{H}_{(k)}} \left(\frac{1}{4}\right)^{|k-k'|} \sum_{k''\in\mathcal{H}_{(k')}} \left(\frac{1}{4}\right)^{|k'-k''|} \\
&\quad +\tau_m^2 \sum_{k'\in\mathcal{H}_{(k)}} \left(\frac{1}{4}\right)^{|k-k'|} \sum_{k''\in\mathcal{F}_{k'}} \left(\frac{1}{4}\right)^{|k'-k''|}
\end{aligned} \tag{5.7}$$

for k' upstream of k.

Then using the recursive relationship in $X(k)$, it is seen that

$$\begin{aligned}
\mathrm{Var}(X(k)) &= \frac{1}{4}\mathrm{Var}(X(u_1))+\frac{1}{4}\mathrm{Var}(X(u_2))+\mathrm{Var}(M(k))+\mathrm{Var}(V(k)) \\
&\quad +2\mathrm{Cov}(\frac{1}{2}X(u_1),M(k))+2\mathrm{Cov}(\frac{1}{2}X(u_2),M(k)) \\
&= \frac{1}{4}\mathrm{Var}(X(u_1))+\frac{1}{4}\mathrm{Var}(X(u_2))+\mathrm{Var}(M(k))+\mathrm{Var}(V(k)) \\
&\quad +\frac{1}{2}\mathrm{Cov}(X(u_1),M(u_1))+\frac{1}{2}\mathrm{Cov}(X(u_2),M(u_2))
\end{aligned}$$

$$
\begin{aligned}
&\vdots \\
&= \sum_{k' \in \mathcal{H}_{(k]}} \left(\frac{1}{4}\right)^{|k'-k|} \mathrm{Var}(V(k')) + \sum_{k' \in \mathcal{F}_k} \left(\frac{1}{4}\right)^{|k'-k|} \mathrm{Var}(X(k')) \\
&\quad + \sum_{k' \in \mathcal{H}_{(k]}} \left(\frac{1}{4}\right)^{|k'-k|} \mathrm{Var}(M(k')) + \sum_{k' \in \mathcal{H}_{(k)}} \left(\frac{1}{4}\right)^{|k'-k|} \frac{1}{2}\mathrm{Cov}(M(k'), X(k')) \\
&= \sigma_v^2 \sum_{k' \in \mathcal{H}_{(k]}} \left(\frac{1}{4}\right)^{|k'-k|} + \tau_x^2 \sum_{k' \in \mathcal{F}_k} \left(\frac{1}{4}\right)^{|k'-k|} \\
&\quad + \sum_{k' \in \mathcal{H}_{(k]}} \left(\frac{1}{4}\right)^{|k'-k|} \mathrm{Var}(M(k')) \\
&\quad + \frac{1}{2} \sum_{k' \in \mathcal{H}_{(k)}} \left(\frac{1}{4}\right)^{|k'-k|-1} \mathrm{Cov}(M(k'), X(k')) \qquad (5.8)
\end{aligned}
$$

where (5.3) and (5.7) can be used to obtain the final expression for the variance.

The covariance between $X(k)$ and any upstream $X(k')$ is derived to be

$$
\begin{aligned}
\mathrm{Cov}(X(k), X(k')) &= \left(\frac{1}{2}\right)^{|k'-k|} \mathrm{Var}(X(k')) + \sum_{k'' \in \mathcal{I}_{(k',k]}} \left(\frac{1}{2}\right)^{|k''-k|} \mathrm{Cov}(X(k'), M(k'')) \\
&= \left(\frac{1}{2}\right)^{|k'-k|} \mathrm{Var}(X(k')) \\
&\quad + \sum_{k'' \in \mathcal{I}_{(k',k]}} \left(\frac{1}{2}\right)^{|k''-k|} \left(\frac{1}{2}\right)^{|k''-k'|} \mathrm{Cov}(X(k'), M(k')) \\
&= \left(\frac{1}{2}\right)^{|k'-k|} \mathrm{Var}(X(k')) \\
&\quad + |k'-k| \left(\frac{1}{2}\right)^{|k'-k|} \mathrm{Cov}(X(k'), M(k'))
\end{aligned}
$$

where $k'' \in \mathcal{I}_{(k',k]}$ is the set of all intermediary reaches between k and k', including k but excluding k'.

Remark 5.1.3 As with the RW+N, second moments can be derived for models based on sums rather than averages. The second moments for $M(k)$ were established in Remark 5.1.1. Using the recursive relationships in the model,

$$
\mathrm{Cov}(X(k), M(k)) = \mathrm{Cov}(X(u_1) + X(u_2) + M(k) + V(k), M(k))
$$

$$
\begin{aligned}
&= \sum_{k' \in \mathcal{H}_{(k]}} \mathrm{Cov}(M(k'), M(k)) \\
&= \sum_{k' \in \mathcal{H}_{(k]}} \mathrm{Var}(M(k')) \\
&= \sum_{k' \in \mathcal{H}_{(k]}} n_{1,k'}\sigma_u^2 + n_{2,k'}\tau_m^2
\end{aligned}
$$

where $n_{1,k'}$ and $n_{2,k'}$ are the sizes of $\mathcal{H}_{(k']}$ and $\mathcal{F}_{k'}$. The second moments of $X(k)$ are defined by

$$
\begin{aligned}
\mathrm{Var}(X(k)) &= \mathrm{Var}(X(u_1)) + \mathrm{Var}(X(u_2)) + \mathrm{Var}(M(k)) + \mathrm{Var}(V(k)) \\
&\quad +2\mathrm{Cov}(X(u_1), M(k)) + 2\mathrm{Cov}(X(u_2), M(k)) \\
&= \mathrm{Var}(X(u_1)) + \mathrm{Var}(X(u_2)) + \mathrm{Var}(M(k)) + \mathrm{Var}(V(k)) \\
&\quad +2\mathrm{Cov}(X(u_1), M(u_2)) + 2\mathrm{Cov}(X(u_2), M(u_2)) \\
&\quad \vdots \\
&= \sum_{k' \in \mathcal{F}_k} \tau_x^2 + \sum_{k' \in \mathcal{H}_{(k]}} \sigma_v^2 \\
&\quad + \sum_{k' \in \mathcal{H}_{(k]}} \mathrm{Var}(M(k')) + 2 \sum_{k' \in \mathcal{H}_{(k)}} \mathrm{Cov}(X(k'), M(k'))
\end{aligned}
$$

with covariances

$$
\mathrm{Cov}(X(k), X(k')) = \mathrm{Var}(X(k')) + |k - k'|\mathrm{Cov}(X(k'), M(k'))
$$

where $|k - k'|$ is the number of confluences between k and k'.

5.1.2.1 Other variations

Other variations of a LLT model may involve combinations of the sums and averages previously discussed. Second moments can be derived to compare the behavior between a variety of local linear trend models, as one may be better suited for a particular process under study. We do not pursue thise variations further in this dissertation.

5.1.2.2 Second differences in LLT

Using the difference operator defined in §2.3.1, we see that

$$\begin{aligned}
\nabla^2 Y(k) &= Y(k) - \frac{2}{2}\left(Y_{11}(k) + Y_{12}(k)\right) + \frac{1}{4}\left(Y_{21}(k) + Y_{22}(k) + Y_{23}(k) + Y_{24}(k)\right) \\
&= W(k) - \left(W_{11}(k) + W_{12}(k)\right) + \frac{1}{4}\left(W_{21}(k) + W_{22}(k) + W_{23}(k) + W_{24}(k)\right) \\
&\quad + X(k) - \frac{2}{2}\left(X_{11}(k) + X_{12}(k)\right) + \frac{1}{4}\left(X_{21}(k) + X_{22}(k) + X_{23}(k) + X_{24}(k)\right) \\
&= W(k) - \left(W_{11}(k) + W_{12}(k)\right) + \frac{1}{4}\left(W_{21}(k) + W_{22}(k) + W_{23}(k) + W_{24}(k)\right) \\
&\quad + V(k) - \frac{1}{2}\left(V_{11}(k) + V_{12}(k)\right) + M(k) - \frac{1}{2}\left(B_{11}(k) + B_{12}(k)\right) \\
&= W(k) - \left(W_{11}(k) + W_{12}(k)\right) + \frac{1}{4}\left(W_{21}(k) + W_{22}(k) + W_{23}(k) + W_{24}(k)\right) \\
&\quad + V(k) - \frac{1}{2}\left(V_{11}(k) + V_{12}(k)\right) + U(k)
\end{aligned}$$

where the lag subscripts clearly indicate dependence up to and including lag two. The differenced process is then 2-correlated, with second moments defined by

$$\begin{aligned}
\gamma(0) &= \sigma_u^2 + \frac{3}{2}\sigma_v^2 + \frac{13}{4}\sigma_w^2 \\
\gamma(1) &= -\frac{1}{2}\sigma_v^2 - \frac{6}{4}\sigma_w^2 \\
\gamma(2) &= \frac{1}{4}\sigma_w^2
\end{aligned}$$

so that results from §4.5 can be used to show the existence of an MA(2) representation for a white noise sequence $\{Z^*(k)\}$ with variance σ^2 such that

$$\nabla^2 Y(k) = \theta(B) Z^*(k),$$

with $\theta(B)$ a second order polynomial in terms of the backshift operator B. A twice differenced $X(k)$ results in

$$\begin{aligned}
\nabla^2 X(k) &= X(k) - \frac{2}{2}\left(X_{11}(k) + X_{12}(k)\right) + \frac{1}{4}\left(X_{21}(k) + X_{22}(k) + X_{23}(k) + X_{24}(k)\right) \\
&= X(k) - \frac{1}{2}\left(X_{11}(k) + X_{12}(k)\right) \\
&\quad - \frac{1}{2}\left(X_{11}(k) + X_{12}(k)\right) - \frac{1}{2}\left(X_{21}(k) + X_{22}(k) + X_{23}(k) + X_{24}(k)\right) \\
&= M(k) + V(k) - \frac{1}{2}\left(M_{11}(k) + V_{11}(k) + M_{12}(k) + V_{12}(k)\right)
\end{aligned}$$

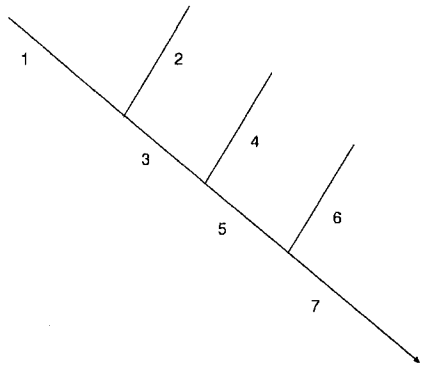


Figure 5.1: Running second order segment similar to Rock Creek.

$$= V(k) - \frac{1}{2}\left(V_{11}(k) + V_{12}(k)\right) + U(k)$$

which can be seen to be a 1-correlated process with variances and covariances defined by

$$\begin{aligned} \gamma(0) &= \frac{3}{2}\sigma_v^2 + \sigma_u^2 \\ \gamma(1) &= \frac{-1}{2}\sigma_v^2 \end{aligned}$$

with $\gamma(h) = 0$ for all $h \geq 2$.

5.1.2.3 Exact Smoother

As discussed in Chapter 2, the use of diffuse priors and the exact Kalman Smoother is specific to the geometric structure of the stream network as well as the underlying stochastic trend model of interest. We consider a special case of the LLT model where $V(k) = 0$ for all k to show that the structure of Rock Creek will not allow the use of the diffuse prior since first order reaches continually enter throughout the network. We use Figure 5.1 as an example as it illustrates the consequences of diffuse initialization over a tree structure when the underlying model does not allow for it.

By a direct extension of the random walk, the diffuse prior on first order k under the LLT model leads to

$$\Omega^p_{\infty,k} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Omega^p_{\star,k} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$\Delta_{\infty,k} = 1 \qquad \Delta_{\star,k} = \sigma_w^2$$

from which we obtain

$$\begin{aligned} X^p(k)^{(0)} &= \mathbf{0} & \mathrm{v}^p(k)^{(0)} &= Y(k) \\ X^p(k)^{(1)} &= \mathbf{0} & \mathrm{v}^p(k)^{(1)} &= 0 \end{aligned}$$

through the recursions defined in §2.5.2. The first filtering step yields

$$X^f(k)^{(0)} = \begin{bmatrix} Y(k) \\ 0 \end{bmatrix} \qquad \Omega^f_{\infty,k} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Omega^f_{\star,k} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & 0 \end{bmatrix}$$

when $Y(k)$ is observed. To begin the recursions, apply this to reaches 1 and 2 in Figure 5.1 to obtain a prediction for reach 3 as well as its variance conditioned on these two observations. This results in

$$X^p(3)^{(0)} = \begin{bmatrix} \frac{Y(1)+Y(2)}{2} \\ 0 \end{bmatrix} \qquad \mathrm{v}^p(3)^{(0)} = Y(k)$$

and

$$\Omega^p_{\infty,k} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \qquad \Omega^p_{\star,k} = \begin{bmatrix} 1/2\sigma_w^2 + \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 \end{bmatrix}$$

Furthermore, we have

$$\Delta_{\infty,3} = 1/2 \qquad \Delta_{\star,3} = \sigma_w^2/2 + \sigma_u^2$$

where we see the innovations variance is still infinite as $\kappa \to \infty$ since $\Delta_{\infty,3}$ is non-zero. Straightforward calculations will show that

$$\Omega^f_{\infty,3} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

to be used in the variance of the next downstream prediction. However, we see that the filtered variance for reach 4 is the same that for reaches 1 and 2, and that $\Omega^f_{\infty,4} \neq 0$, where it is easy to see through the recursions that the diffuseness is reintroduced into the prediction variance associated with reach 5. The consequence is that now $\Omega^p_{\infty,5} \neq 0$ and the innovations variance once again tends to infinity. Although it can be shown that $\Omega^f_{\infty,5}$ is the zero matrix, diffuseness will always be introduced when a first order reach enters the network, a situation prevalent in the Rock Creek tree structure.

With the infinite prediction and innovation variances, the likelihood is greatly influenced as the innovation terms will tend to zero as $\kappa \to \infty$ whereas the determinant terms will become infinite. Hence, it is clear that we can not apply a diffuse prior with the LLT model on Rock Creek.

5.1.3 Discrete Smoothing Spline

Suppose we had a univariate series through the stream network, and wished to approximate the series with a smooth function $\mu(k)$. A standard approach (Durbin and Koopman, 2001) would be to determine $\mu(k)$ that minimizes the criterion function

$$\sum_{k=1}^{n} (y(k) - \mu(k))^2 + \lambda \sum_{k=1}^{n} \left(\nabla^2 \mu(k)\right)^2 \tag{5.9}$$

with respect to $\mu(k)$, where $\lambda > 0$ and $\nabla^2\mu(k)$ is twice differenced $\mu(k)$. The minimization involves a trade-off between goodness of fit to the data and smoothness of the fit, where the degree of smoothness is determined by λ. The solution to this minimization is a smoothing spline, or a penalized least squares estimator. For large λ, more emphasis is placed on the $\nabla^2\mu(k)$ terms in the minimization which results in a smoother spline fit. For small λ, more weight is placed on the squared residuals, and $\mu(k)$ will be close to $y(k)$ resulting in a better fit but less smooth $\mu(k)$.

For a traditional spline smoother in discrete time, the smoothness parameter λ can be determined either through cross validation techniques or through a function of variance components in a local linear trend model. Since the local linear trend is a state-space model, the Kalman Smoother can be used to obtain both the spline estimates as well as confidence bounds at each point.

The existence of a spline smoother on a stream network can be seen through the special case of the local linear trend model defined in §5.1.2 but with $V(k) = 0$. In this model there are two variance components: variation due to measurement error and that due to fluctuation of the underlying process, specifically the slope.

Suppose we wish to smooth the $Y(k)$ series by estimating $X(k)$ by $\hat{X}(k) = E(X(k)|\mathbf{Y})$. Under the Gaussian error structure, $X^s(k) = E(X(k)|\mathbf{Y})$ is readily obtained via the Kalman recursions once initialization requirements are specified, where we will see that the degree of smoothness is controlled by set values of σ_w^2 and σ_u^2. Alternatively, these smoothed estimates can be found through calculus via maximization of the posterior density $p(\mathbf{X}|\mathbf{Y})$. Since $X(k)$ and $Y(k)$ are Gaussian, this posterior mode is also the posterior mean, which is also the conditional mean.

The joint density of $(\mathbf{X}, \mathbf{Y})$ is $p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})$. The density $p(\mathbf{Y}|\mathbf{X})$ is straightforward since $Y(k)|\mathbf{X}$ are independent $N(X(k), \sigma_w^2)$. However, writing out $p(\mathbf{X})$ requires some assumptions pertaining to the first order reaches. In order to establish second differences, state and slope terms of imaginary upstream reaches are assumed to be known. We can then write the density of $\mathbf{X}$ as a product of conditional densities $p(X(k)|X(k'), k' \in \mathcal{U}_{(k)})$ where the conditioning is on upstream $X(k')$. We start with the first order reaches and progress downstream with flow.

Under this special case of the local linear trend model, these conditional densities are

$$p(X(k)|X(k'), k' \in \mathcal{U}_{(k)}) \sim N\left(\frac{1}{2}(X(u_1) + X(u_2)) + \frac{1}{2}(M(u_1) + M(u_2)), \sigma_u^2\right) \quad (5.10)$$

where the slope terms $M(u_i)$ are known since $X(u_i)$ and both of its parents are used in the conditioning. From this we see that the conditional means depend on the grandparents through the slope terms. It is clear from (5.10) that the likelihood can be fully expressed if $X(u_i)$ and M_{u_i}, for $i = 1, 2$ are assumed known for any first order reach.

Alternatively, we can use the backshift operator on $X(k)$ to see that

$$\begin{aligned}
\nabla^2 X(k) &= (1 - 2B + B^2)X(k) \\
&= X(k) - 2(\frac{1}{2})\,(X_{11}(k) + X_{12}(k)) + \frac{1}{4}\,(X_{21}(k) + \ldots + X_{24}(k)) \\
&= M(k) - \frac{1}{2}X_{11}(k) + \frac{1}{4}(X_{21}(k) + X_{22}(k)) \\
&\quad -\frac{1}{2}X_{12}(k) + \frac{1}{4}(X_{23}(k) + X_{24}(k)) \\
&= M(k) - \frac{1}{2}(X_{11}(k) - \frac{1}{2}(X_{21}(k) + X_{22}(k)) - \frac{1}{2}(X_{12}(k) \\
&\quad -\frac{1}{2}(X_{23}(k) + X_{24}(k)) \\
&= M(k) - \frac{1}{2}M(u_1) - \frac{1}{2}M(u_2) \\
&= U(k)
\end{aligned}$$

thus yielding a sequence of independent Normal random variables. From this we find that the density of the twice differenced $X(k)$ is equivalent to expression of the independent conditional densities, conditioned on upstream reaches.

The posterior distribution is defined by $p(\mathbf{X}|\mathbf{Y}) = p(\mathbf{X}, \mathbf{Y})/p(\mathbf{Y})$. Since

$$\log p(\mathbf{X}|\mathbf{Y}) = \log p(\mathbf{X}, \mathbf{Y}) - \log p(\mathbf{Y})$$

it is clear that maximizing $\log p(\mathbf{X}|\mathbf{Y})$ with respect to $\mathbf{X}$ is equivalent to maximizing $\log p(\mathbf{X}, \mathbf{Y})$ with respect to $\mathbf{X}$ since $p(\mathbf{Y})$ is independent of $\mathbf{X}$. For the model considered here, maximizing $\log p(\mathbf{Y}, \mathbf{X})$ is equivalent to maximizing

$$-\frac{1}{2\sigma_w^2}\sum_{k=1}^{n}(Y(k) - X(k))^2 - \frac{1}{2\sigma_u^2}\sum_{k=1}^{n}(\nabla^2 X(k))^2.$$

with respect to each $X(k)$.

When $\lambda = \sigma_w^2/\sigma_u^2$, this is equivalent to minimizing (5.9). The solution to the normal equations from $\log p(\mathbf{X}, \mathbf{Y})$ with respect to $\mathbf{X}$ is the mode of $p(\mathbf{X}|\mathbf{Y})$, which is $\hat{\mathbf{X}} = E(\mathbf{X}|\mathbf{Y})$ under normality assumptions. Thus, the Maximum Likelihood estimator of $\mathbf{X}$ is the same as that from the Kalman Smoother. Since the estimator for $\mathbf{X}$ from the Kalman Smoother is also a penalized least squares estimator, we obtain a spline smoother in discrete space through the special case of a Local Linear Trend.

Example 5.1.1 Montgomery County, MD

We now return to the data from Montgomery County, Maryland. The data of interest here are water chemistry collected at each of the stations. Variables such as dissolved oxygen, PH, air temperature, and water temperature are available for monitoring stations throughout the county at a variety of time points. For simplicity, we selected only a small drainage to work with, that of Upper and Lower Rock Creek. Many reaches have multiple monitoring stations that may have collected data during summer and winter months since the mid 1990s. The geometry of this specific watershed and distribution of observation stations can be seen in Figure 4.14. There are 39 reaches total.

We specifically model dissolved oxygen content, as the levels of dissolved oxygen and consequently that of dissolved carbon dioxide indicate metabolism of organisms (Angelier, 2003, p.97). There are many ways in which oxygen is introduced as well as depleted from water causing concentrations to be highly variable. In polluted waters with high organic matter content such as litter and decomposing leaves, aerobic bacteria consume large amounts of dissolved oxygen (Allan, 1995, p.24), reducing its availability for other processes in which it is a necessity.

Much of the dissolved oxygen in water comes from the atmosphere. The solubility of oxygen is primarily a function of temperature and partial pressure of oxygen in its gaseous state (Allan, 1995, p.24). The influence of temperature is obvious by

its seasonal and daily changes in its levels. Typical levels range from 16.64 mg/l at 0°C to 7.54 mg/l at 30°C (Angelier, 2003, p.97). Gases dissolve better in cold water. The warmer the water temperature, the lower the dissolved oxygen content.

The amount of oxygen dissolved is also a function of the water surface interface, with turbulence favoring gas exchange. Low water flow decreases dissolved oxygen by decreasing the amount of air/water mixing that occurs in rapids and waterfalls.

Photosynthesis and respiration are the two important biological processes that alter the levels of dissolved oxygen. Algae and rooted aquatic plants deliver oxygen to water through photosynthesis. Conversely, fish, invertebrates, plants, and aerobic bacteria all require oxygen for respiration, thus leading to a reduction in dissolved oxygen. Moreover, the presence of pollution increases the demands from respiration further depleting the water of oxygen (Allan, 1995, p.24).

Our interest is in modeling dissolved oxygen as a smooth function throughout the stream network. In efforts to reduce some of the natural variation, only data for summer months was used. A non-parametric smoothing spline is attractive in that a smooth function can be obtained without having to assume some parametric model for the large number of covariates that influence dissolved oxygen levels. The spline is especially appealing here since many of the influential covariates such as water and air temperature may not be available at every reach.

The spline function allows a researcher to assess the behavior of dissolved oxygen throughout the network, and reaches in which problematic levels can be easily identified. The goal is to obtain two separate smooth functions for each the main channels above the confluence of the two second order segments shown in Figures 4.14 and 5.2. Using a LLT model, we obtain a smoothing spline to accomplish this goal.

A closed form expression exists for maximum likelihood estimators for the mean of the distribution of initial states. The two variance components associated with

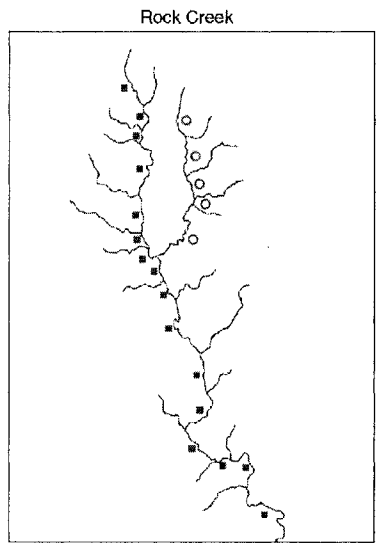
Table 5.1: Maximum Likelihood Estimates for Rock Creek

$\hat{x_0}$= 7.48	$\hat{b}_0$= -2.29	$\hat{\sigma}_w^2$= 4.24
$\hat{\tau}_x^2$= 0.08	$\hat{\tau}_b^2$= 9.84	$\hat{\sigma}_u^2$= 3.23
		$\hat{\lambda}$= 1.312

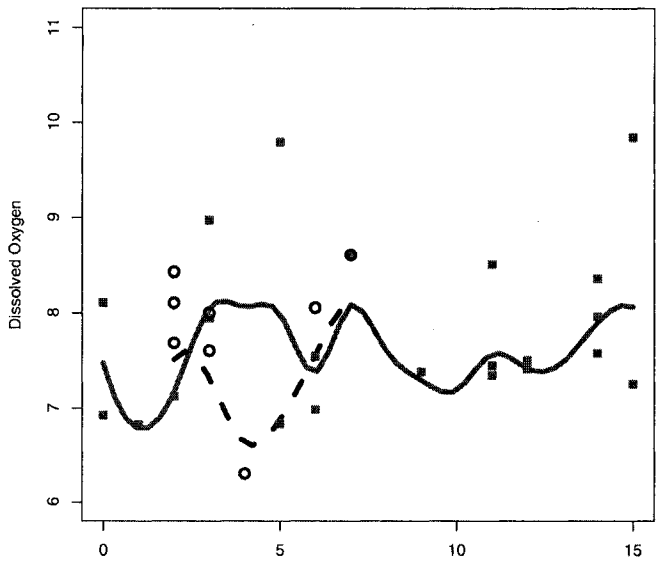
the initial states as well as the two variance components of the model are estimated using numerical optimization. Parameter estimates are shown in Table 5.1. The resulting spline is found in Figure 5.2 where we also see in this figure that the squares indicate a sequence of points down one main channel whereas the circles represent the points from the other. Note that there is only one sequence of points the third order segment.

With 18 first order reaches having data, the estimated mean of the initial starting point, x_0, is reasonable, and close to a simple mean of the observations associated with the first order reaches. The initial slope term is of less interest than its estimated variance. We see a fairly large variance associated with these initial slopes, which in turn allows the smoothed estimates to be close to the observed data, consistent with an undersmoothed function.

The degree of smoothness was chosen via MLE for the model variance components. Variation due to multiple observations on each reach is apparent in Figure 5.2. With summer averages (over years), the resulting variance estimate for σ_w^2 is close to zero, and the spline appears to be undersmoothed.



(a) Identification of channels on Rock Creek



(b) Spline

Figure 5.2: Dissolved oxygen versus lag for Upper and Lower Rock Creek. As seen in the reference plot, the squares represent data along one main channel while circles represent data from a second channel that merges with the first. Solid line is a smoothing spline fit to the data along the first main channel whereas the dashed line is the spline fit to the data from the second channel. Smoothing parameter determined by MLE's of variance components in LLT model 5.1.2 but with $V(k) = 0$.

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this dissertation, we have adapted a number of modeling techniques from time series to fit the tree-like structure of a stream network, primarily through development of a state-space representation and the Kalman Recursions. Dependence on a stream network is modeled as a function of flow rather than time.

We have defined a number of both stationary and nonstationary models for a stream network, and determined a state-space form for each. With the state-space form, we have defined the Kalman recursions for a tree structure. Although these recursions were developed in a Gaussian context, the best linear prediction properties of the Kalman Filter and Smoother still hold even in the non-Gaussian case, where predictions are based on projections rather than conditional expectations.

Two variations of a Kalman Filter and Smoother have been derived for the stream network. For cases when initial conditions are unknown, we have been able to estimate these conditions through numerical optimization with appropriate model assumptions for first order reaches. Diffuse initialization was also considered, where we have shown its limited application because of the nature of a stream network. Unlike the case of a time series which has a single starting point, our estimation of initial conditions seems to be a reasonable approach on a stream network, since there are many starting points and our model assumptions for first order reaches allow for it.

The Kalman Filter applied to the stream network also allows for determination of an exact likelihood, even when missing data are present. This Gaussian likelihood

can then be used to obtain Maximum Likelihood Estimators for model parameters. A concentrated likelihood may also be obtained in some cases, and numerical techniques for optimization can be used to achieve estimates for remaining unknown quantities. Estimates can then be used in conjunction with the Kalman Smoother to obtain final predicted values for the state of interest.

We have also applied an EM algorithm as an approximation when the likelihood is complicated by missing data. Simulation has shown that in the example considered, estimates from this EM algorithm converged to those obtained through the exact likelihood. A simplified, or adjusted, EM algorithm can also be used as an approximation, although simulation has shown this to increase the magnitude of the bias and consequently the mean squared error in parameter estimation. Nonetheless, we have applied these commonly used techniques for parameter estimation on a stream network with the use of the Kalman recursions developed in Chapter 2.

The class of ARMA(p, q) models have been defined for the stream network, and concepts of causality and invertibility developed in terms of mean square convergence. We found that these concepts are well defined through functions of the autoregressive and moving average polynomials for the ARMA(p, q) model. The polynomial roots of these "scaled" polynomials can also be used to formulate autocovariance functions for the model of interest. We are also able to calculate a sample autocorrelation function for univariate data, where a sampling approach to construction of this function can be used when multiple observations exist.

In general, simulations have shown that processes on a stream network behave similar to those in time series. A variety of forms of dependence have been shown with the ARMA models considered. We have found that a non-zero autoregressive parameter tends to have more influence on the shape of the ACF than the moving average parameter. Some of the models even describe alternating signs of dependence. While the physical existence of such correlation for a stream network may

seem unrealistic, applicability of these models extends to many other tree structures where processes can evolve from many root nodes to one single terminal node.

We have used the sample ACVF to construct method of moment estimators, which can then be used as initial values in numerical optimization in methods for likelihood based estimators of model parameters. We have shown that the method of moments estimators do not guarantee the non-negative definite constraints for covariances, and the Yule-Walker equations do not assure sample estimates of a causal process as seen with the AR(1). Likewise, the sample lag-1 autocorrelation for an MA(1) is often well outside the boundary of $1/2\sqrt{2}$ defined by invertibility. This is often the case when the data are representative of a process with parameters near the invertible boundaries. Causality and invertibility can be guaranteed in some cases with Maximum Likelihood Estimators through transformations of model parameters.

We have found the sample ACF to be a useful diagnostic in model selection. Realizations can be generated and corresponding sample ACFs calculated and compared with the observed ACF as an ad-hoc method to assess if the observed data are representative of a process defined by the model fit.

We have taken fish habitat data on Rock Creek and fit the AR(1), MA(1), ARMA(1, 1) and AR(2) models to a measure of instream cover. Empirical white noise bounds in conjunction with model fit ACF identify the MA(1) or MA(2) as possible models. This is supported by a parametric bootstrap which indicated a significant Moving Average parameter for the MA(1). The instream cover was then fit to a linear model using reach specific covariates such as land use, elevation, gradient, and cover type. Non-significance of the MA(1) parameter indicated that the spatial autocorrelation in instream cover was removed by regressing on appropriate covariates. This analysis showed that possible autocorrelation can sometimes easily be removed. Formal tests for Goodness-of-fit and model selection were not performed, although they are left as an open area of research.

We have defined a class of stochastic trend models for a stream network. Similar to that of time series, these models can be used to describe a process in terms of a trend component, an advantage when appropriate covariates may not be obtainable. A differencing operator can be applied to these trend models to achieve stationarity, but the resulting tree structure and number of observations used are greatly reduced. Moreover, these models can be redefined such that the state is defined in terms of sums or weighted averages, rather than simple averages of upstream parents.

With a special case of the Local Linear Trend model described in Chapter 5, we are able to define a smoothing spline on a stream network. We have shown that in the Gaussian case, the predictions resulting from the Kalman Smoother are also penalized least squares predictions, thus generating a spline on the network. The smoothing parameter can be estimated through maximum likelihood estimators of the variance parameters, and the resulting spline obtained by executing the Kalman recursions as a function of the estimated variance components.

We have demonstrated the applicability of the smoothing spline with data from Rock Creek. It is well known that dissolved oxygen is influenced by many covariates such as turbidity, depth, and water temperature. Yet it is clear that covariates such as these are not realistically obtainable, making the smoothing spline an attractive tool to describe not only trends, but potentially problematic areas in the network with unusual levels of dissolved oxygen.

We believe we have only begun to develop modeling tools for tree structures such as a stream network through state-space models. Many of the ideas for modeling and forecasting from time series have network analogues, thus application on a stream network. The use of data on a small section of Rock Creek has not only demonstrated applicability, but also has brought to attention a number of practical areas that deserve more attention such as limits due to network geometry and missing values. With that, there is no shortage of ways to enhance and expand upon the tools developed in this dissertation.

6.1 Future Work

The extension of the state-space models in time series is an intuitive approach to modeling and forecasting on a stream network. Although we have touched upon defining a large class of stationary models as well as some stochastic trend models, there are many areas that warrant consideration of future research. We have taken the approach of modeling a discrete process, which we see as a building block for other models.

Throughout this dissertation, we assumed that the state remains constant across the entire reach. A logical step forward seems to investigate the possibility of allowing the state to change within a reach using information such as distance from a confluence or distance between monitoring stations.

We have developed method-of-moment estimators as well as those of Maximum Likelihood for the models considered. Although we have achieved these estimates under specified constraints, reliability of these estimates remains as an open area of study. We have been able to determine approximate mean square errors through simulation, but rely on a parametric bootstrap for such estimates when modeling data for Rock Creek. Extension of a non-parametric bootstrap utilized in time series seems impractical for the tree structures here. Investigation of the asymptotic properties of these Maximum Likelihood estimators should also be considered.

Although techniques for parameter estimation are developed, there is an immediate need for diagnostic tools for data exploration and model validation. We generate data under the fitted model and compare to that observed as one technique. Other methods to assess dependence and stationarity deserve attention.

There are many other types of dependence commonly used in spatial statistics, such as the Matern class (Stein, 1999, p.31) and other, more complex spatial-temporal covariance structures. It only seems reasonable to consider network analogues of these models.

We have developed the recursions in the Gaussian context, although we believe this is not a restriction. The recursions can similarly be interpreted as projections to achieve optimal linear predictors. There may be cases where a linear model may be inappropriate, or inadequately describe the behavior of the data.

Although the main focus of this dissertation has been model development and application to stream networks, there are many other areas in which a process can start at many initial nodes and proceed to merge until a single final node is reached. Examples range from root systems in plants and the cardiovascular system in animals, to genetic traits that are common from generation to generation. We believe that our modeling techniques developed here have direct application in other areas.

In conclusion, we believed we have only touched upon the development of models for stream networks analogous to those commonly used in time series. There are simply many more areas that have yet to be addressed. We believe we have established a base which can be used to extend the concepts developed and serve as a starting point for many remaining areas, and hope to pursue these in the future.

APPENDIX I: Notation

The following table summarizes much of the notation used throughout this dissertation, and is intended for ease in reference.

$\mathcal{U}_{(k)}$	Set of reaches upstream of k
$\mathcal{U}_{(k]}$	Reach k and the set of reaches upstream of k
$\mathcal{D}_{(k)}$	Set of reaches downstream of k
$\mathcal{D}_{[k)}$	Reach k and the set of reaches downstream of k
$\mathcal{S}_i$	Set of all i^{th} order segments
$\mathcal{I}_{(k',k]}$	Set of (intermediary) reaches between k' and k, excluding reach k' but including reach k
$\mathcal{F}_k$	Set of first order reaches upstream of k
$\mathcal{F}$	Set of first order reaches
$\mathcal{H}_{(k]}$	Reach k and all higher order reaches that are upstream of k
$\mathcal{H}$	Set of Higher order reaches
$\mathcal{B}_j$	Set of reaches in the j^{th} basin
$\mathbf{X}_{i,j}(k)$	Vector associated with j^{th} reach at lag i from reach k
$\mathcal{X}_{(k]}$	$\mathbf{X}(k) \cup \{\mathbf{X}_{i,j}(k)\}_{i=1...\infty,j=1...2^i}$
n_{ij}	Number of reaches in i^{th} order segment j
$\|k-k'\|$	Number of confluences between k and k'. Also referred to as *lag*.
$\|\|k-k'\|\|$	Distance between locations k and k'
m_s	Highest Strahler order of all reaches in the network
$(\cdot)^p$	Associated with initial prediction
$(\cdot)^f$	Associated with filtering
$(\cdot)^s$	Associated with smoothing

Bibliography

Allan, J. D. (1995). *Stream Ecology: Structure and function of running waters.* Chapman & Hall.

Angelier, E. (2003). *Ecology of Streams and Rivers.* Science Publishers, Inc.

Basseville, M., Benveniste, A., Chou, K., Golden, S., Nikoukhah, R., and Willsky, A. (1992). Modeling and estimation of multiresolution stochastic processes. *IEEE Transactions on Information Theory*, 38:766–784.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer-Verlad New York Inc.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting.* Springer-Verlad New York Inc.

Chou, K. C., Willsky, A. S., and Nikoukhah, R. (1994). Multiscale systems, Kalman filters, and Riccati equations. *IEEE Transactions on Automatic Control*, 39:479–492.

Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley and Sons, New York.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:127–150.

De Jong, P. (1988). The likelihood for a state space model. *Biometrika*, 75:165–169.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Seried B (Methodological)*, 39:1–39.

Dent, C. and Grimm, N. (1999). Spatial heterogeneity of stream water nutrient concentrations over successive time. *Ecology*, 80:2283–2298.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State-Space Methods.* Oxford University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press.

Hocking, R. R. (1996). *Methods and Application of Linear Models.* John Wiley & Sons, Inc.

Horton, R. E. (1945). Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Bull. Geological Soc. America*, 56:275–370.

Huang, H. and Cressie, N. (2001). Multiscale graphical modeling in space: Application to command and control. *Spatial Statistics: Methodological Aspects and Some Applications, Springer Lecture Notes in Statistics*, 159. M. Moore, ed. Springer-Verlag, New York.

Huang, H., Cressie, N., and Gabrosek, J. (2002). Fast, resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics*, 11:63–88.

Huang, H. C. (1997). *Spatial modeling using graphical Markov models and wavelets.* PhD thesis, Iowa State University.

Johannesson, G. and Cressie, N. (2004). Finding large scale spatial trends in massive, global, environmental datasets. *Environmetrics*, 15:1–44.

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

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for nonstationary time series. *Journal of the American Statistical Association*, 92:1630–1638.

Koopman, S. J. and Durbin, J. (2003). Filtering and smoothing of state vector for diffuse state-space models. *Journal of Time Series Analysis*, 24:85–98.

Lange, K. (1998). *Numerical Analysis for Statisticians.* Springer-Verlag, New York.

Mielke, P. and Berry, K. (2001). *Permutation Methods : A Distance Function Approach.* Springer, New York.

Monestiez, P., Bailly, J., Lagacherie, P., and Voltz, M. (2005). Geostatistical modelling of spatial processes on directed trees: Application to fluvisol extent. *Geoderma*, 128:179–191.

Peterson, E., Merton, A., Theobald, D., and Urquhart, N. (2006). Patterns of spatial autocorrelation in stream water chemistry. *Environmental Monitoring and Assessment*, 121:571–596.

Peterson, E. and Urquhart, N. (2006). Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: A case study in maryland. *Environmental Monitoring and Assessment*, 121:615–638.

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

Strahler, A. N. (1957). Quantitative analysis of watershed geomorphology. *Transactions of the American Geophysical Union*, 38:913–920.

Tzeng, S., Huang, H., Cressie, N., and Gabrosek, J. (2005). A fast, optimal spatial-prediction method for massive datasets. *Journal of American Statistical Association*, 100(472):1343–1357.

Ver Hoef, J. M., Peterson, E. E., and Theobald, D. M. (2006). Some new spatial statistical models for stream networks. *Environmental and Ecological Statistics*, 14.

Verghese, G. and Kailath, T. (1979). A further note on backwards Markovian models. *IEEE Transactions on Information Theory*, 25:121–124.