Smoothing through State-Space Models for Stream Networks

William J. Coar
DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY

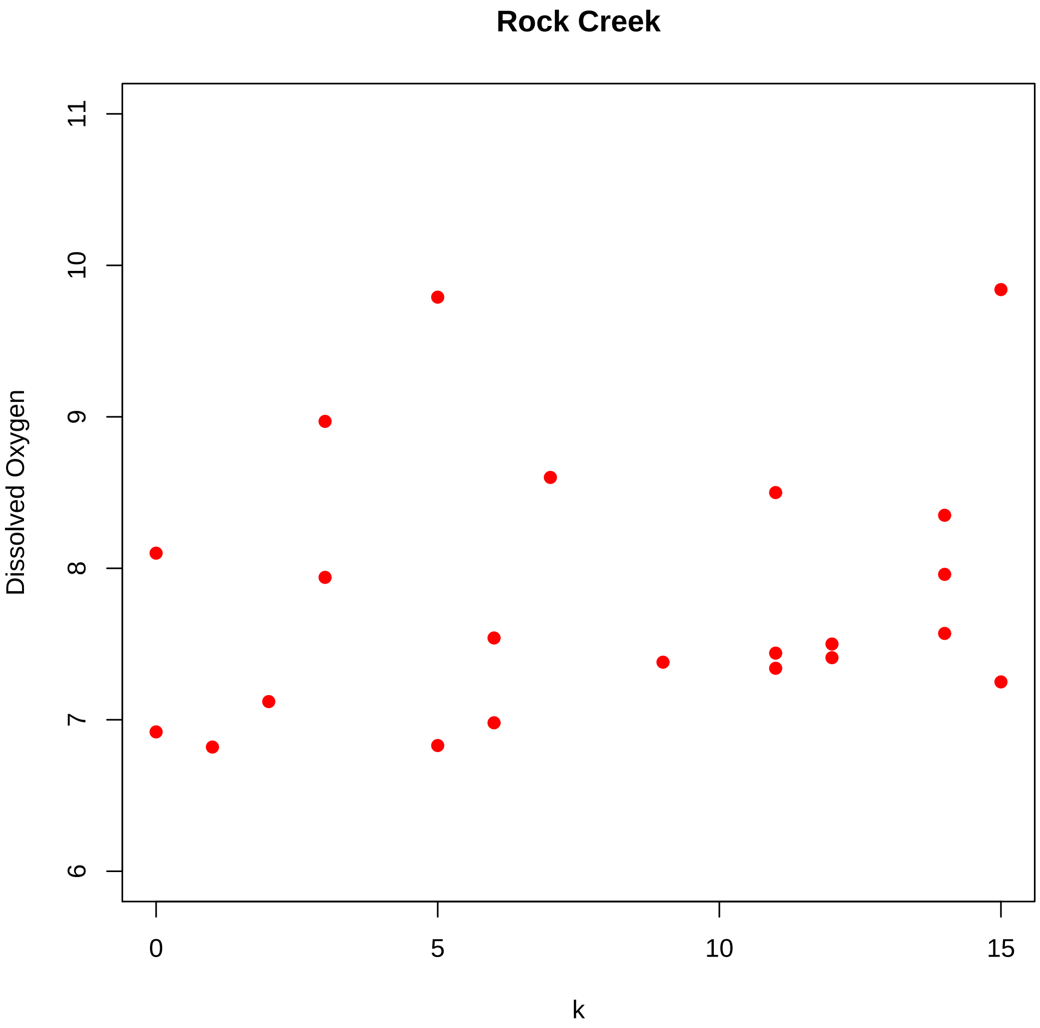
August 7, 2006

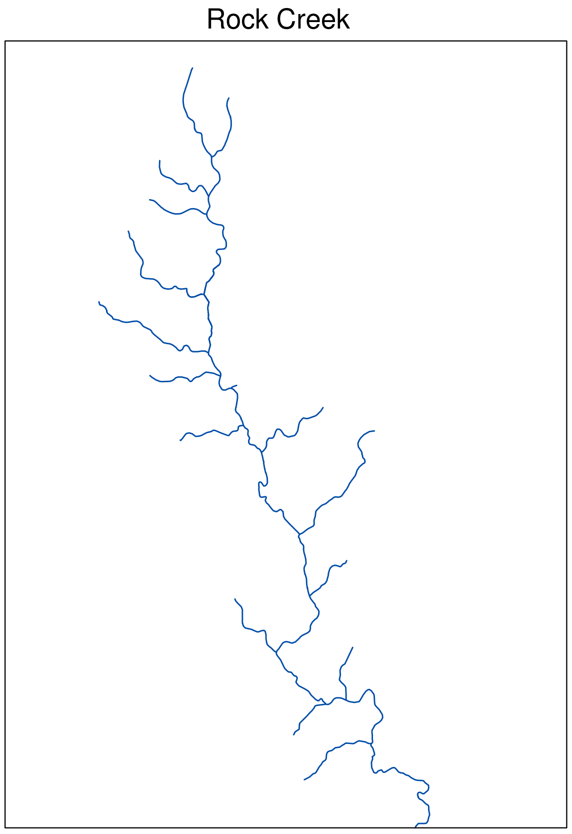
Joint work with F. Jay Breidt, Colorado State University.

Outline

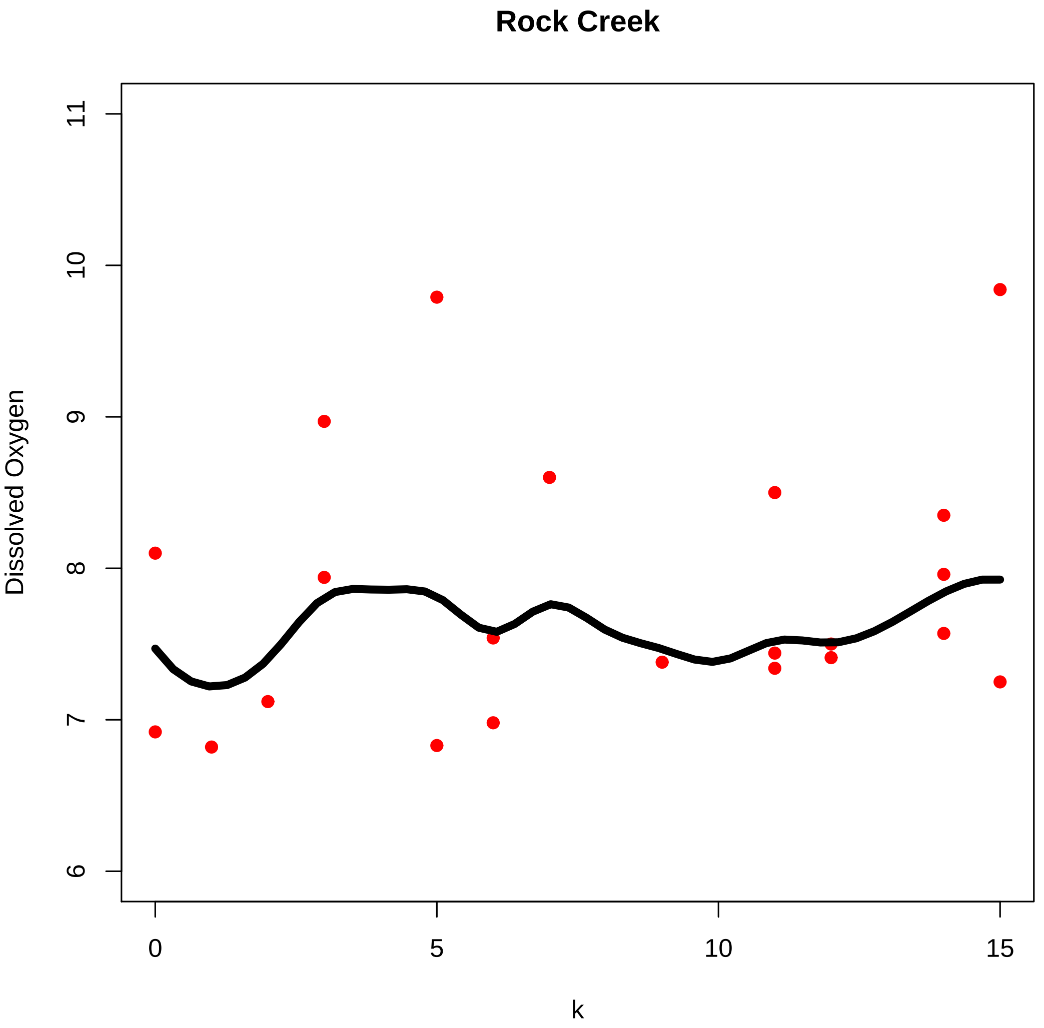
1. A smoothing problem $\longrightarrow$ traditional spline smoother
2. Our goal
3. Stream networks
   (a) Local linear trend
   (b) State-space representation
4. Kalman recursions
5. Connection to a discrete spline smoother
6. Numerical example

One path on a stream network

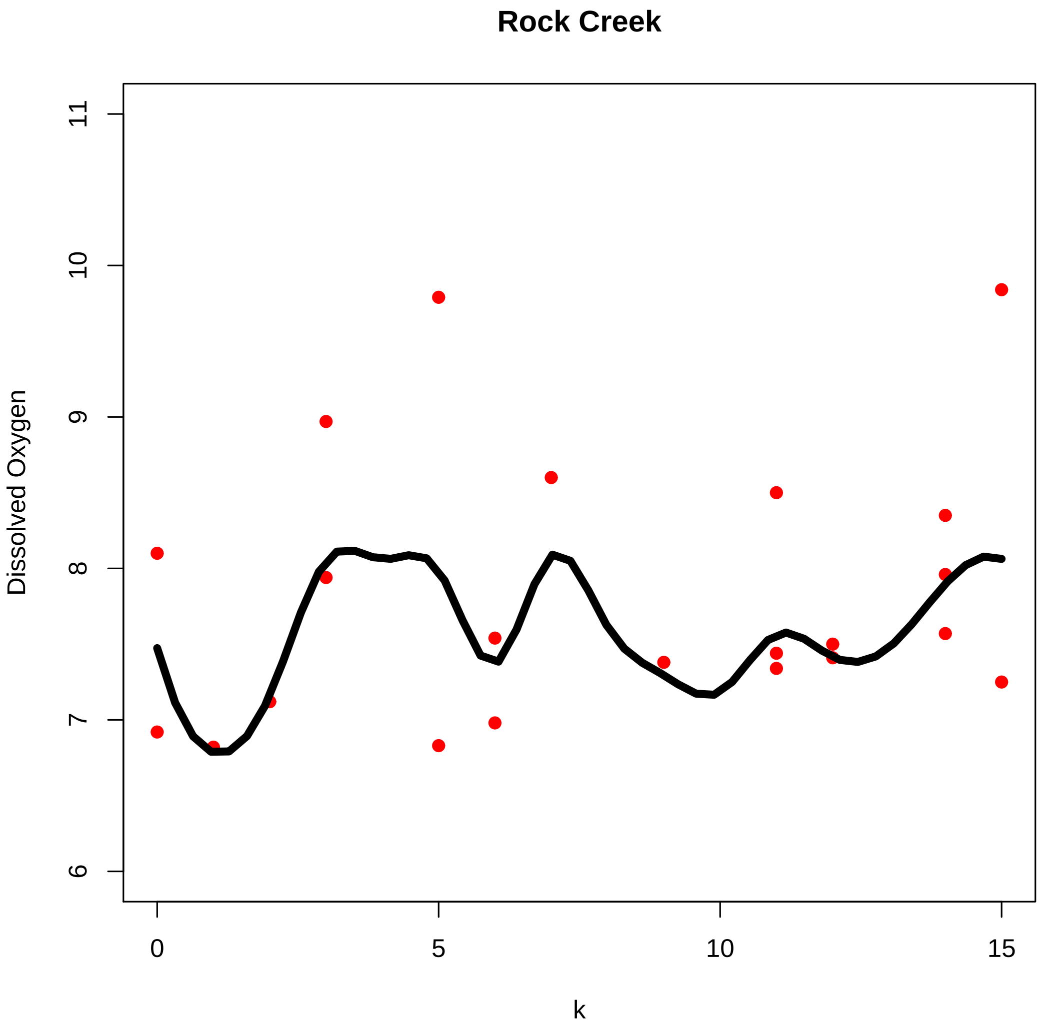




How much to smooth?



Maybe a *better* smooth



Traditional spline smoother in discrete time

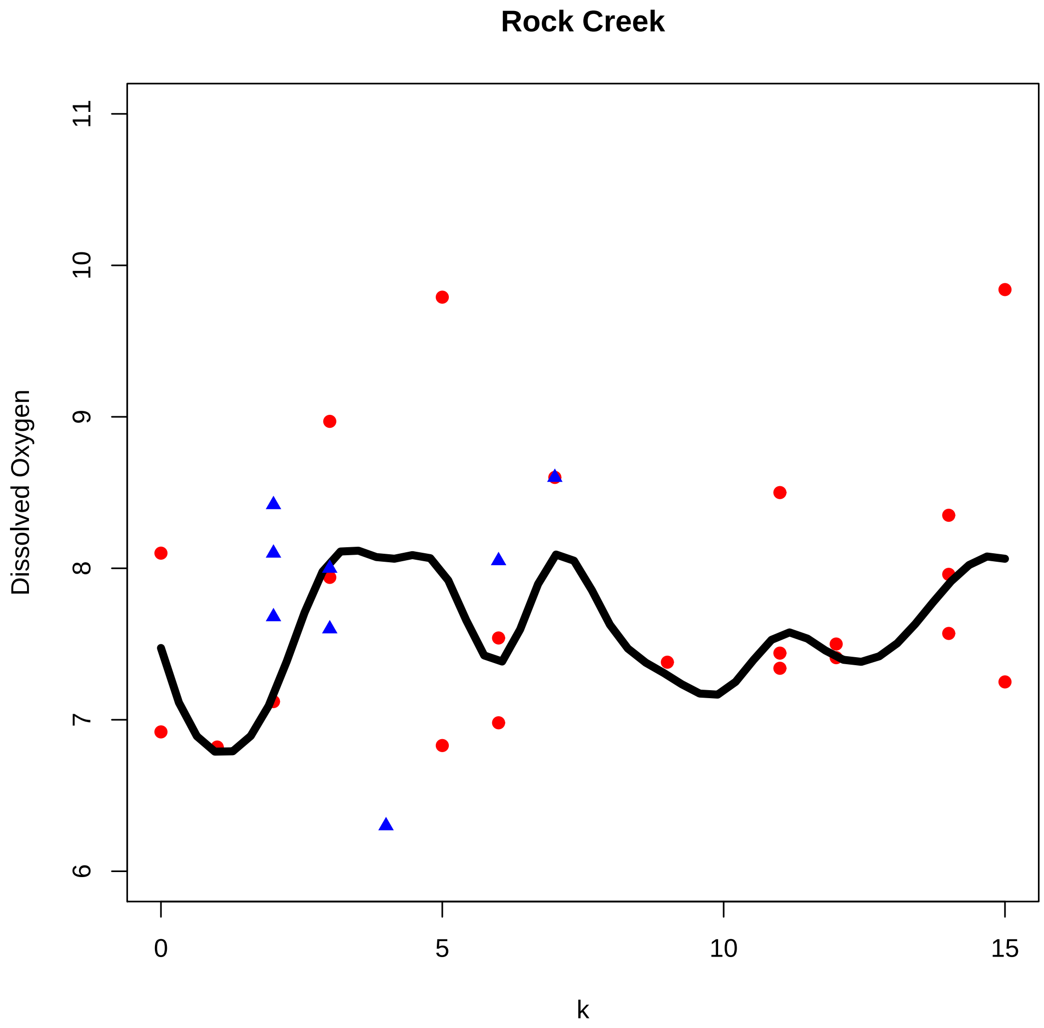
Smooth $\mu(t)$ that minimizes a penalized least squares criterion function

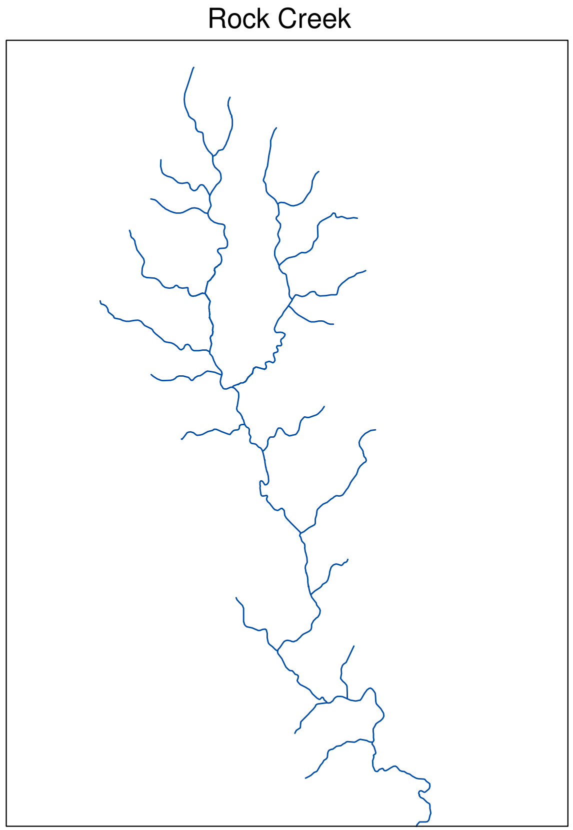
$$\sum_{t=1}^{n} (y(t) - \mu(t))^2 + \lambda \sum_{t=1}^{n} \left(\nabla^2 \mu(t)\right)^2$$

where $\nabla^2 \mu(t)$ is twice differenced $\mu(t)$.

- Trade off between goodness-of-fit and smoothness
- Smoothness determined by λ
  1. Cross-validation
  2. Function of variance components in Local Linear Trend
- Local Linear Trend (LLT) is a state-space model
- Spline obtained as the Kalman smooth using this state-space representation

A new problem → two paths merge



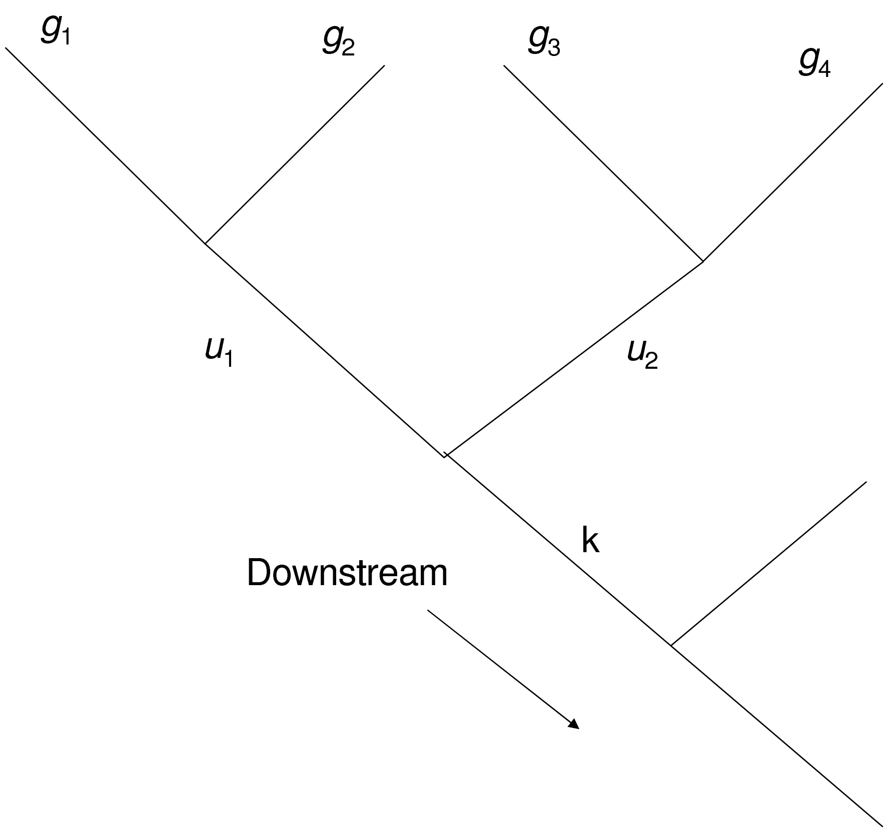


Our Goals

Adapt time series methods to smooth the network

- Define a Local Linear Trend model
- Determine its state-space representation
- Implement Kalman recursions
- Construct a “spline” smoother on the network

Diagram of a stream network



- Reach characterized by (Strahler) order
- Except for first order reaches, each reach k has two parents u_1 and u_2
- Some reaches have grandparents
- Process on reach k depends on the states at each parent
- Natural time-like ordering $\rightarrow$ downstream flow
- Merging with each step downstream
- For simplicity, assume equally spaced discrete locations

State-space model and Local Linear Trend

State-space model on network:

Observation equation $Y(k)$ = Linear function of (unobservable) state + noise
Process equation $\mathbf{X}(k)$ = Linear function of the state at upstream reaches + noise

Local Linear Trend model:

$$\begin{aligned} Y(k) &= \text{Level at reach } k \text{ plus noise} \\ X(k) &= \text{Average levels at parent reaches + slope component} \\ B(k) &= \text{Average slopes at parent reaches + noise} \end{aligned}$$

- Variation associated with measurement error
- Variation due to fluctuation in underlying random process (slope)

State-space model on network:

$$Y(k) = G_k \mathbf{X}(k) + W(k) \qquad (\mathbf{Y}_t = G_t \mathbf{X}_t + \mathbf{W}_t)$$
$$\mathbf{X}(k) = F_{k,u_1} \mathbf{X}(u_1) + F_{k,u_2} \mathbf{X}(u_2) + \mathbf{V}(k) \quad (\mathbf{X}_t = F_t \mathbf{X}_{t-1} + \mathbf{V}_{t-1})$$

Local Linear Trend model:

$$Y(k) = X(k) + W(k) \qquad (\mathbf{Y}_t = \mathbf{X}_t + \mathbf{W}_t)$$
$$X(k) = \tfrac{1}{2}\left(X(u_1) + X(u_2)\right) + B(k) \quad (X_t = X_{t-1} + B_{t-1}$$
$$B(k) = \tfrac{1}{2}\left(B(u_1) + B(u_2)\right) + U(k) \quad (B_t = B_{t-1} + U_{t-1})$$

with state-space components

$$\mathbf{X}(k) = \begin{bmatrix} X(k) \\ B(k) \end{bmatrix} \quad F_{k,u_i} = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1/2 \end{bmatrix} \quad \mathbf{V}(k) = \begin{bmatrix} U(k) \\ U(k) \end{bmatrix}$$

Smoothing via Kalman recursions

Downstream *predict, filter, predict, filter...*

Given upstream information, predict via

$$\begin{aligned} \mathbf{X}^p(k) &= \text{Prediction based solely on upstream observations} \\ \Omega_k^p &= \text{Variance of this prediction} \end{aligned}$$

Filter once observation is obtained

$$\begin{aligned} \mathbf{X}^f(k) &= \text{Update original prediction with new information from } \mathbf{Y}(k) \\ \Omega_k^f &= \text{Variance of this updated prediction} \end{aligned}$$

Upstream *smooth*

$$\begin{bmatrix} \mathbf{X}^s(u_1) \\ \mathbf{X}^s(u_2) \end{bmatrix} = \text{Update the filtered values with information from downstream reaches}$$

RESULT: Expectation based on all observed information.

Smoothing via Kalman recursions

Downstream *predict, filter, predict, filter...*

Given upstream information, predict via

$$\begin{aligned} \mathbf{X}^p(k) &= F_{k,u_1}\mathbf{X}^f(u_1) + F_{k,u_2}\mathbf{X}^f(u_2) \\ \Omega_k^p &= F_{k,u_1}\Omega_{u_1}^f F_{k,u_1}^T + F_{k,u_2}\Omega_{u_2}^f F_{k,u_2}^T + Q_t, \end{aligned}$$

Filter once observation is obtained

$$\begin{aligned} \mathbf{X}^f(k) &= \mathbf{X}^p(k) + \Omega_k^p G_k^T \Delta_k^{-1}\left(\mathbf{Y}(k) - G_k\mathbf{X}^p(k)\right) \\ \Omega_k^f &= \Omega_k^p - \Omega_k^p G_k^T \Delta_k^{-1} G_k \Omega_k^p. \end{aligned}$$

where $\Delta_k = G_k\Omega_k^p G_k^T + R_k$.

Upstream *smooth*

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

where $\Theta(u_i,k) = \Omega_{u_i}^f F_{k,u_i}^T (\Omega_k^p)^{-1}$.

RESULT: Smoothed estimates $E(\mathbf{X}|\mathbf{Y})$.

Conditional/Posterior mean

For Gaussian case, conditional/posterior mean is conditional/posterior mode

Conditional/Posterior mode: most probable $\mathbf{X}$ given $\mathbf{Y}$, the mode of $p(\mathbf{X}|\mathbf{Y})$

Another way to get conditional/posterior mode:

- Maximize $p(\mathbf{X}|\mathbf{Y})$ using calculus
- Equivalent to maximizing

$$-\frac{1}{2\sigma_w^2}\sum_{k=1}^{n}(Y(k)-X(k))^2-\frac{1}{2\sigma_u^2}\sum_{k=1}^{n}(\nabla^2 X(k))^2.$$

where $\nabla^2 X(k) = U(k)$

Conditional/Posterior mode is Penalized Least Squares

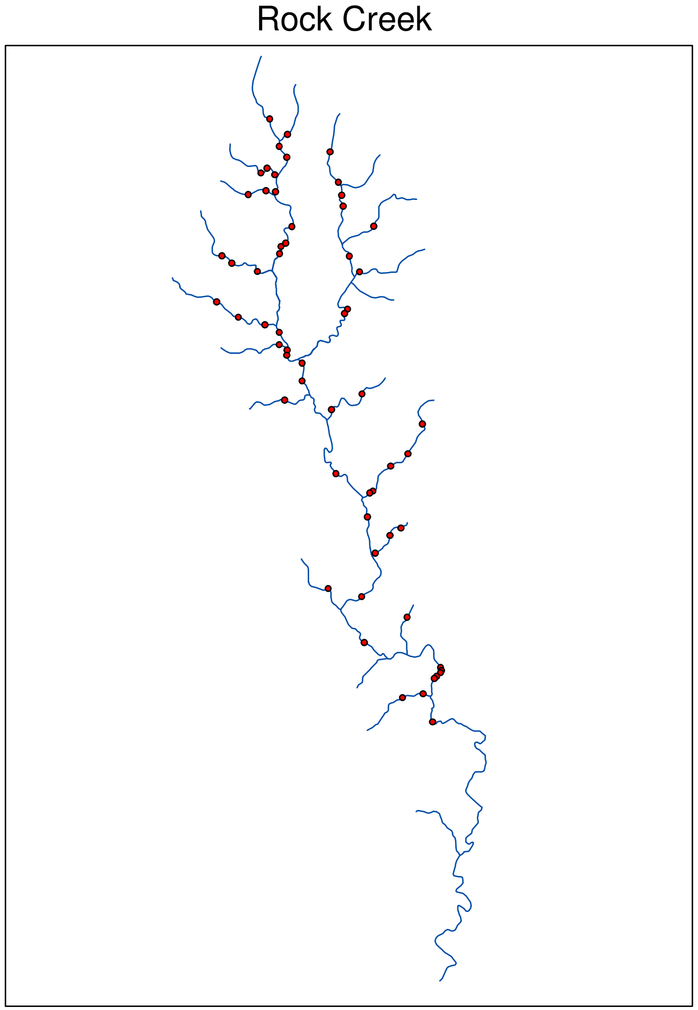
- or equivalently, minimize

$$\sum_{k=1}^{n}(Y(k) - X(k))^2 + \frac{\sigma_w^2}{\sigma_u^2}\sum_{k=1}^{n}(\nabla^2 X(k))^2.$$

  (as was used for traditional spline)

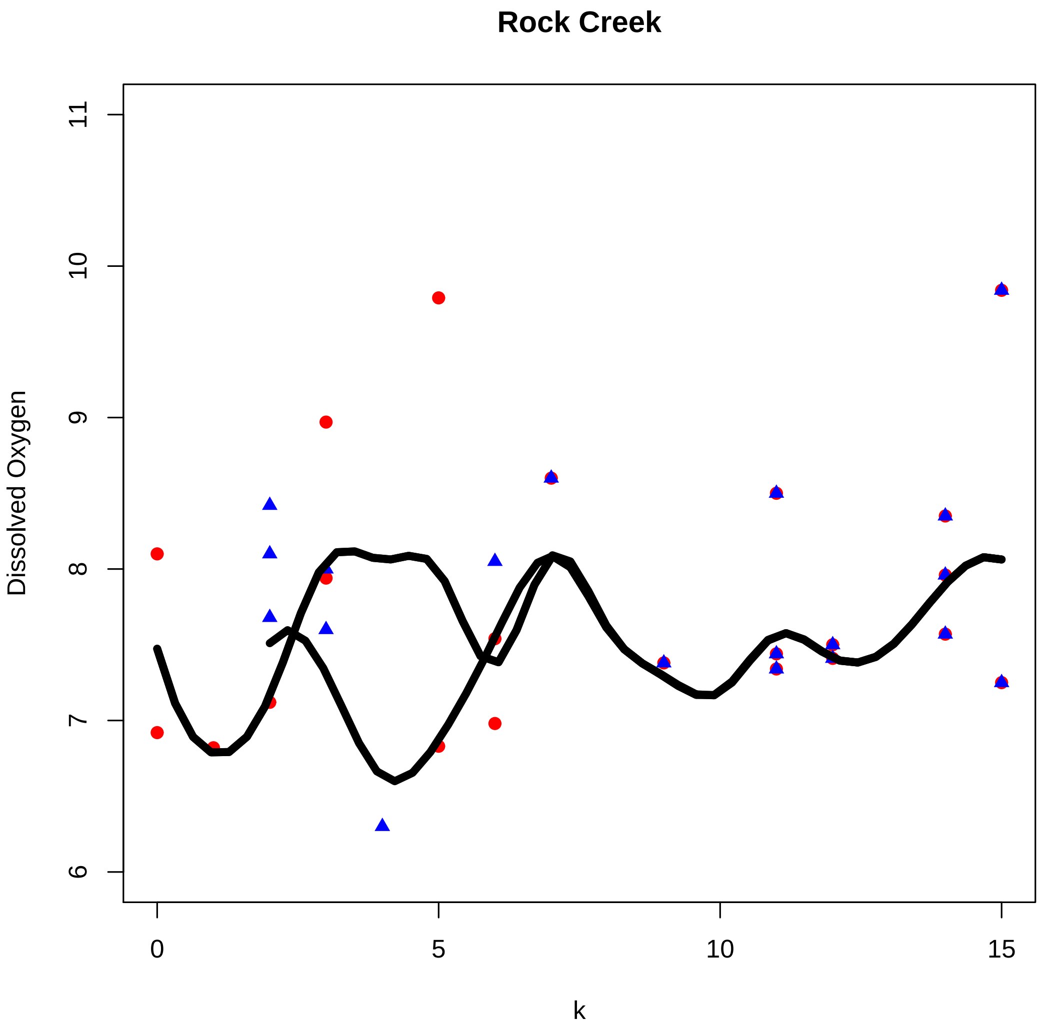
- This defines a *spline smoother* on a stream network through LLT
- Obtain estimate of smoothness parameter $\lambda = \frac{\sigma_w^2}{\sigma_u^2}$ by maximum likelihood
- Obtain $E\left[\mathbf{X}|\mathbf{Y}\right]$ via Kalman smoother

Example - Upper and Lower Rock Creek, MD



- Measure dissolved oxygen
- Summer months
- Some reaches with > 1 station
- Some stations with > 1 observation
- For simplicity, multiple obs/reach treated as independent reps

Example 1: $\hat{\lambda} = 1.312$ with estimated initial conditions



Results

Estimation of initial conditions

- For first order k, $\mathbf{X}(k) \overset{i.i.d}{\sim} N\left(\begin{bmatrix} \mu_0 \\ b_0 \end{bmatrix}, \begin{bmatrix} \sigma_\mu^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}\right)$

Mean $\hat{\mu_0} = 7.48$ **Slope** $\hat{b_0} = -2.29$

$\hat{\sigma^2}_\mu = 0.08$ $\hat{\sigma^2}_b = 9.84$

Estimation of smoothing parameter

$\hat{\sigma_w^2} = 4.24$

$\hat{\sigma_u^2} = 3.23$

$\hat{\lambda} = 1.312$

Other considerations

- Addition of seasonal and yearly effects

Further work on State-Space Models

State-space model for stream network:

$$\begin{aligned} Y(k) &= G_k \mathbf{X}(k) + W(k) \\ \mathbf{X}(k) &= F_{k,u_1}\mathbf{X}(u_1) + F_{k,u_2}\mathbf{X}(u_2) + \mathbf{V}(k) \end{aligned}$$

General form is very flexible

- Can be multivariate
- A time component can be added, but process driven by flow
- State matrices are location dependent

Describe a large class of dependencies

- Class of ARMA(p,q) models can be defined
- More general structural models (LLT)

Work in progress

- Adapted state-space to a stream network
- Defined ARMA(p,q) and other structural models on a stream network
- Developed Kalman recursions for this state-space representation
- Likelihood in terms of innovations
- An EM algorithm for missing values
- Application to real data - Montgomery County, MD
  - Network structure from FLoWS
  - Enormous thanks to Stephanie Fitchett, John Norman, and Dave Theobald

The work reported here was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the presenter and STARMAP, the Program he represents. EPA does not endorse any products or commercial services mentioned in this presentation.

This research is funded by
U.S.EPA – Science To Achieve Results (STAR) Program
Cooperative Agreement # CR - 829095