Generalizing the Matrix Normal Distribution

– An application to spatio-temporal data

Ekvall, Karl Oskar [University of Minnesota] and Gray, Brian [US Geological Survey]
Contact: ekvall@umn.edu

Goal

Develop model suitable for estimating trends in and fitting splines to time series of environmental and ecological outcomes from multiple spatial locations that takes into account both spatial and temporal dependence.

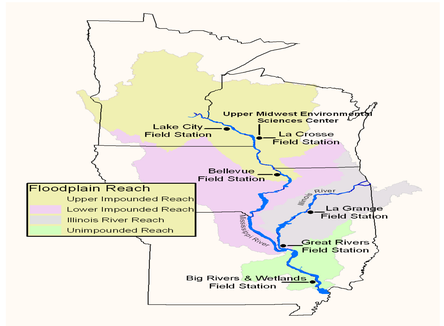
Data

- Water temperature measurements from $\sim$ 20 locations on the Mississippi river
- Sampled quarterly for 20 years
- Each response component is the sample mean of $\sim 5-150$ measurements

Spatial Structure

Locations represent an incomplete crossing of reaches and sampling strata.

Reaches on the Mississippi [Image source: www.umesc.usgs.gov]



Temporal Structure

- May assume data from different years are independent
- There may or may not be dependence between data from different seasons within the same year

$$\underset{r\times c}{\boldsymbol{Y}_t} = \overset{\text{3 Seasons}}{\begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{18,1} & y_{18,2} & y_{18,3} \end{pmatrix}}, t = 1, \dots, n$$

A Data Excerpt

```
## # A tibble: 1,080 x 5
##    location  year season     temp     n
##      <fctr> <int> <fctr>    <dbl> <int>
## 1       1:1  1994     SP 10.12000    25
## 2       1:2  1994     SP 10.62333    30
## 3       1:3  1994     SP 12.32600    50
## 4       1:4  1994     SP 11.08333    30
## 5       2:1  1994     SP 12.18000    25
## 6       2:2  1994     SP 11.82333    30
## 7       2:3  1994     SP 12.95500    60
## 8       2:5  1994     SP 11.69600    25
## 9       2:6  1994     SP 15.41000    10
## 10      3:1  1994     SP 15.85000    30
## # ... with 1,070 more rows
```

The Matrix Normal Distribution

$$\boldsymbol{y}_t := \text{vec}(\boldsymbol{Y}_t) \sim \text{N}_{rc}\left(\text{vec}(\boldsymbol{X}_t\boldsymbol{\beta}), \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\Sigma} = \boldsymbol{V} \otimes \boldsymbol{U}$$

- Restriction: $\text{cov}(y_{i,j}, y_{i',j'}) = \boldsymbol{U}_{i,i'}\boldsymbol{V}_{j,j'}$ [separability of covariance]
- Gain: $\mathcal{O}(r^2 + c^2)$ covariance parameters instead of $\mathcal{O}(r^2c^2)$
- Note: Independence would imply $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_{rc}^2) \in \mathcal{O}(rc)$

Theory: MLE of $\boldsymbol{\Sigma}$ exists and is unique when $n > r/c + c/r + p$, which is proved in [1]. The key observation is that the negative log-likelihood is geodesically convex in $(\boldsymbol{V}, \boldsymbol{U})$ with the geodesics $d_t(\boldsymbol{A}, \boldsymbol{B}) := \boldsymbol{A}^{1/2}(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2})^t\boldsymbol{A}^{1/2}$

Our Generalization

$$\boldsymbol{y}_t := \text{vec}(\boldsymbol{Y}_t) \sim \text{N}_{rc}\left(\text{vec}(\boldsymbol{X}_t\boldsymbol{\beta}), \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\Sigma} = \boldsymbol{C}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{C},$$

where $\boldsymbol{C} = \text{diag}(\sigma_1, \dots, \sigma_{rc})$.

- Restriction: $\text{corr}(y_{i,j}, y_{i',j'}) = \boldsymbol{B}_{i,i'}\boldsymbol{A}_{j,j'}$ [separability of *correlation*]
- Gain: Completely general variance structure
- Still with $\mathcal{O}(r^2 + c^2) \ni \mathcal{O}(r^2 + c^2 + rc)$ parameters

Theory: The negative log-likelihood is convex in each of $\boldsymbol{V}$, $\boldsymbol{U}$ and $\boldsymbol{C}_{ii}, i = 1, \dots, rc$, but not jointly. We do not know whether the negative log-likelihood is geodesically convex under some appropriate choice of geodesics.

Optimization

We propose a blockwise coordinate descent algorithm for maximizing the likelihood.

- In each update, we set an unrestricted gradient to zero and then rescale variables to satisfy the constraint
- All updates are closed form

For example, treating $\boldsymbol{A}$ as a *general* matrix: update $\boldsymbol{A}^{k+1}$ to the solution of $\nabla_A \ell(\boldsymbol{A}, \boldsymbol{B}^k, \boldsymbol{C}^k) = 0$. Then, rescale:

- $\boldsymbol{C}^k \leftarrow \boldsymbol{C}^k(\text{diag}(\boldsymbol{A}^{k+1})^{1/2} \otimes \boldsymbol{I})$
- $\boldsymbol{A}^{k+1} \leftarrow \text{diag}(\boldsymbol{A}^{k+1})^{-1/2}\boldsymbol{A}^{k+1}\,\text{diag}(\boldsymbol{A}^{k+1})^{-1/2}$

Rescaling after updating both $\boldsymbol{A}$ and $\boldsymbol{B}$ and then updating $\boldsymbol{C}$ component by component gives the following algorithm.

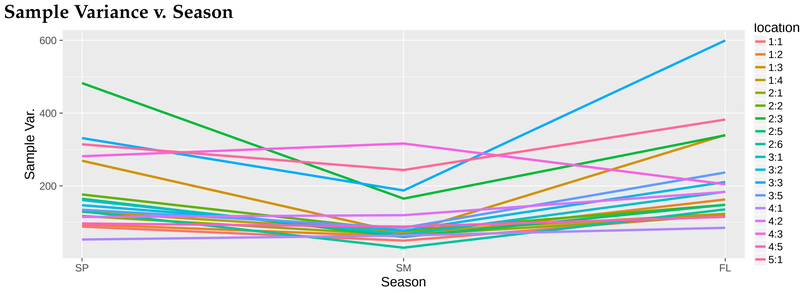
Algorithm: Blockwise coordinate descent

Initialize $\boldsymbol{A}^0, \boldsymbol{B}^0, \boldsymbol{C}^0, k = 0$
repeat
  Set $\boldsymbol{A}^{k+1}$ to the solution of $\nabla_{A^{-1}} \ell(\boldsymbol{A}, \boldsymbol{B}^k, \boldsymbol{C}^k) = 0$
  Set $\boldsymbol{B}^{k+1}$ to the solution of $\nabla_{B^{-1}} \ell(\boldsymbol{A}^{k+1}, \boldsymbol{B}, \boldsymbol{C}^k) = 0$
  Rescale $\boldsymbol{A}^{k+1}, \boldsymbol{B}^{k+1}$ and $\boldsymbol{C}^k$ to satisfy constraints
  for $j = 1, \dots, rc$ **do**
    Set θ_j to the solution of $\nabla_{\theta_j} \ell(\boldsymbol{A}^{k+1}, \boldsymbol{B}^{k+1}, \theta_1^{k+1}, \dots, \theta_{j-1}^{k+1}, \theta_j, \theta_{j+1}^k, \dots, \theta_{rc}^k) = 0$
  end for
  $k \leftarrow k + 1$
until $|\ell^k - \ell^{k-1}| \leq \epsilon$

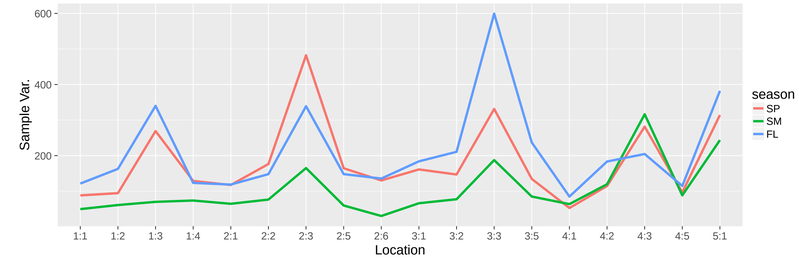
Notation: ℓ^k is the log-likelihood evaluated at the kth iterate of all optimization variables. ∇_Z denotes gradient w.r.t. variable Z. ϵ is a tolerance parameter for terminating the algorithm. We parameterize $\boldsymbol{C}_{i,i} = 1/\theta_i$.

Results

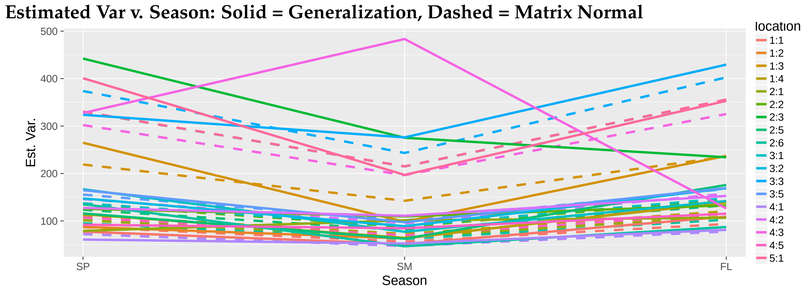
Sample Variance v. Season



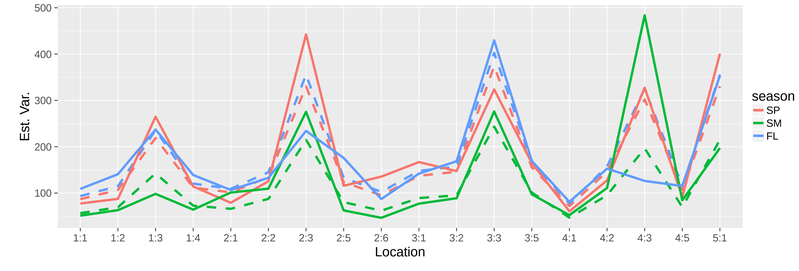
Sample Variance v. Location



Estimated Var v. Season: Solid = Generalization, Dashed = Matrix Normal



Estimated Var v. Location: Solid = Generalization, Dashed = Matrix Normal



References

[1] I. Soloveychik and D. Trushin. Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness. *Journal of Multivariate Analysis*, 149:92 – 113, 2016.