Generalizing the Matrix Normal Distribution

– An application to spatio-temporal data

Karl Oskar Ekvall

September 30, 2016

University of Minnesota

Outline

1. A motivating example
2. The matrix normal distribution
3. A generalization
4. Some (very) preliminary results

US Geological Survey

"Providing the scientific information needed by managers, decision makers, and the public to protect, enhance, and restore the ecosystems in the Upper Mississippi River Basin, the Midwest, and worldwide."

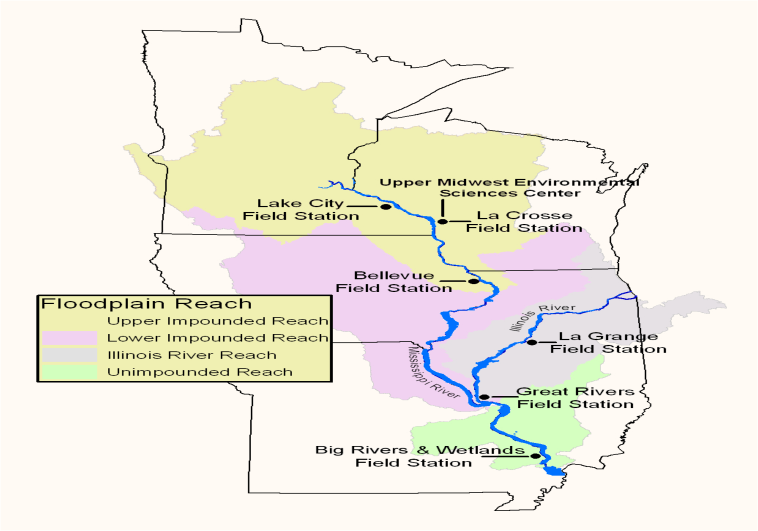
[Source: www.umesc.usgs.gov]

A motivating example

Data

- Average water temperature measurements from ~ 20 locations on the Mississippi river
- Sampled quarterly for ~ 20 years

Spatial structure[1]



[1] Picture from USGS

Temporal structure

$$\boldsymbol{Y}_t = \begin{array}{c} \text{3 Seasons} \\ \begin{pmatrix} y_{1,1} & y_{1,2} & y_{1,3} \\ \vdots & \vdots & \vdots \\ y_{18,1} & y_{18,2} & y_{18,3} \end{pmatrix} \end{array}, t = 1, \ldots, 20$$

Data Excerpt

```
## # A tibble: 1,080 <U+00D7> 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
```

Some characteristics

Based on communication with scientists:

- May assume data from different years are independent

Some characteristics

Based on communication with scientists:

- May assume data from different years are independent
- Within years, data from different seasons may or may not be independent

Some characteristics

Based on communication with scientists:

- May assume data from different years are independent
- Within years, data from different seasons may or may not be independent
- Strong dependence between measurements from different sampling locations taken within the same year

Some characteristics

Based on communication with scientists:

- May assume data from different years are independent
- Within years, data from different seasons may or may not be independent
- Strong dependence between measurements from different sampling locations taken within the same year
- **Quiz:**

Some characteristics

Based on communication with scientists:

- May assume data from different years are independent
- Within years, data from different seasons may or may not be independent
- Strong dependence between measurements from different sampling locations taken within the same year
- **Quiz:** Why may winter be less interesting than other seasons to model?

The matrix normal distribution

A possible parameterization

$$\boldsymbol{y}_t := \mathrm{vec}(\boldsymbol{Y}_t) \overset{iid}{\sim} \mathrm{N}_{54}(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U})$$

A possible parameterization

$$\boldsymbol{y}_t := \text{vec}(\boldsymbol{Y}_t) \overset{iid}{\sim} \mathrm{N}_{54}(\text{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U})$$

- Restriction: $\text{cov}(y_{i,j}, y_{i',j'}) = \boldsymbol{U}_{i,i'} \boldsymbol{V}_{j,j'}$

A possible parameterization

$$\boldsymbol{y}_t := \text{vec}(\boldsymbol{Y}_t) \overset{iid}{\sim} \mathrm{N}_{54}(\text{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U})$$

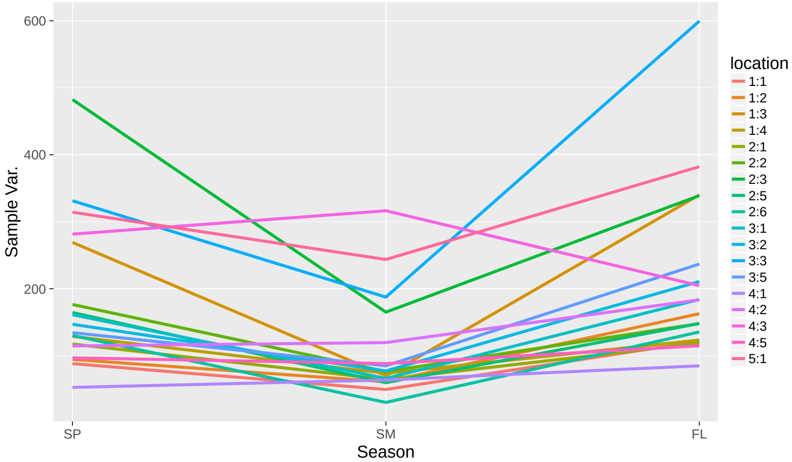
- Restriction: $\text{cov}(y_{i,j}, y_{i',j'}) = \boldsymbol{U}_{i,i'} \boldsymbol{V}_{j,j'}$
- Gain: $3(3+1)/2 + 18(18+1)/2 = 176$ instead of $18(18+1)/2 = 1485$ parameters

A possible parameterization

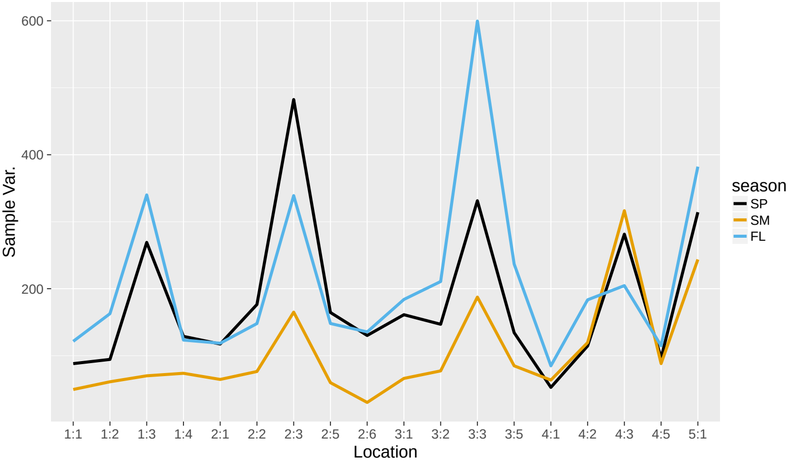
$$\boldsymbol{y}_t := \mathrm{vec}(\boldsymbol{Y}_t) \overset{iid}{\sim} \mathrm{N}_{54}(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U})$$

- Restriction: $\mathrm{cov}(y_{i,j}, y_{i',j'}) = \boldsymbol{U}_{i,i'} \boldsymbol{V}_{j,j'}$
- Gain: $3(3+1)/2 + 18(18+1)/2 = 176$ instead of $18(18+1)/2 = 1485$ parameters
- MLE: Everything is fine as long as $n \geq 18/3 + 3/18 + 1 \approx 8$ [Soloveychik and Trushin, 2016]

Assumptions: All locations the same?



Assumptions: All seasons the same?



A generalization

A more general parameterization

$$\underset{54\times 54}{\boldsymbol{\Sigma}} = \underset{54\times 54}{\boldsymbol{C}}\big(\underset{3\times 3}{\boldsymbol{A}} \otimes \underset{18\times 18}{\boldsymbol{B}}\big)\underset{54\times 54}{\boldsymbol{C}}$$

where $\boldsymbol{C} = \mathrm{diag}(1/\theta_1, \ldots, 1/\theta_{54})$ and $\boldsymbol{A}, \boldsymbol{B}$ are correlation matrices.

A more general parameterization

$$\underset{54\times 54}{\boldsymbol{\Sigma}} = \underset{54\times 54}{\boldsymbol{C}}\big(\underset{3\times 3}{\boldsymbol{A}} \otimes \underset{18\times 18}{\boldsymbol{B}}\big)\underset{54\times 54}{\boldsymbol{C}}$$

where $\boldsymbol{C} = \mathrm{diag}(1/\theta_1, \ldots, 1/\theta_{54})$ and $\boldsymbol{A}, \boldsymbol{B}$ are correlation matrices.

Properties:

- Complete variance heterogeneity

A more general parameterization

$$\underset{54\times54}{\boldsymbol{\Sigma}} = \underset{54\times54}{\boldsymbol{C}}\left(\underset{3\times3}{\boldsymbol{A}} \otimes \underset{18\times18}{\boldsymbol{B}}\right)\underset{54\times54}{\boldsymbol{C}}$$

where $\boldsymbol{C} = \mathrm{diag}(1/\theta_1, \ldots, 1/\theta_{54})$ and $\boldsymbol{A}, \boldsymbol{B}$ are correlation matrices.

Properties:

- Complete variance heterogeneity
- Same correlation structure as matrix normal

A more general parameterization

$$\underset{54\times54}{\boldsymbol{\Sigma}} = \underset{54\times54}{\boldsymbol{C}}\left(\underset{3\times3}{\boldsymbol{A}} \otimes \underset{18\times18}{\boldsymbol{B}}\right)\underset{54\times54}{\boldsymbol{C}}$$

where $\boldsymbol{C} = \mathrm{diag}(1/\theta_1, \ldots, 1/\theta_{54})$ and $\boldsymbol{A}, \boldsymbol{B}$ are correlation matrices.

Properties:

- Complete variance heterogeneity
- Same correlation structure as matrix normal
- Requires $rc - r - c + 1 = 54 - 18 - 3 + 1 = 34$ more parameters

A more general parameterization

$$\underset{54\times 54}{\boldsymbol{\Sigma}} = \underset{54\times 54}{\boldsymbol{C}}\left(\underset{3\times 3}{\boldsymbol{A}} \otimes \underset{18\times 18}{\boldsymbol{B}}\right)\underset{54\times 54}{\boldsymbol{C}}$$

where $\boldsymbol{C} = \mathrm{diag}(1/\theta_1, \ldots, 1/\theta_{54})$ and $\boldsymbol{A}, \boldsymbol{B}$ are correlation matrices.

Properties:

- Complete variance heterogeneity
- Same correlation structure as matrix normal
- Requires $rc - r - c + 1 = 54 - 18 - 3 + 1 = 34$ more parameters
- Still $\mathcal{O}(r^2 + c^2)$, as $(r, c) \to (\infty, \infty)$, compared to $\mathcal{O}(r^2c^2)$ for a general structure

Maximum likelihood

- The likelihood is not convex in general but could still have unique global maximum

Maximum likelihood

- The likelihood is not convex in general but could still have unique global maximum
- The negative log likelihood for the matrix normal is *geodesically convex*; it's not obvious whether this is also true for our model

Maximum likelihood

- The likelihood is not convex in general but could still have unique global maximum
- The negative log likelihood for the matrix normal is *geodesically convex*; it's not obvious whether this is also true for our model
- We propose a blockwise coordinate descent algorithm

Maximum likelihood

- The likelihood is not convex in general but could still have unique global maximum
- The negative log likelihood for the matrix normal is *geodesically convex*; it's not obvious whether this is also true for our model
- We propose a blockwise coordinate descent algorithm
- Every update is convex and in closed form

Algorithm

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 $j = 1, \dots, m$ **do**

    Set θ_j to the solution of $\nabla_{\theta_j} \ell(\boldsymbol{A}^{k+1}, \boldsymbol{B}^{k+1}, \theta_1^{k+1}, \dots, \theta_{j-1}^{k+1}, \theta_j, \theta_{j+1}^k, \dots, \theta_m^k) = 0$

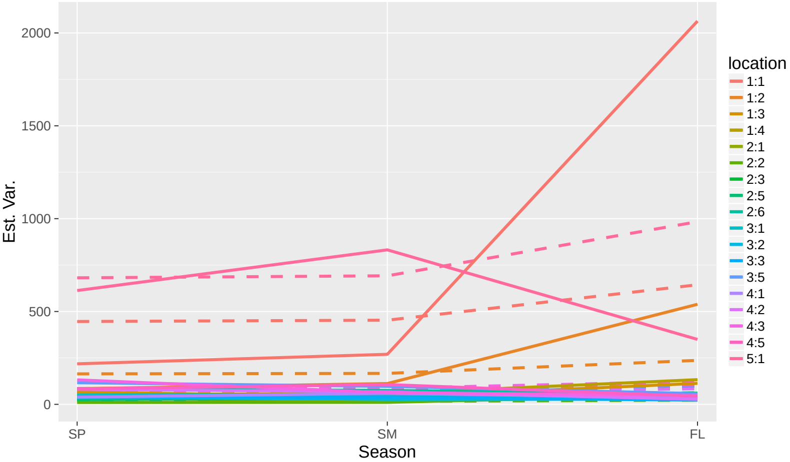
  end for

  $k \leftarrow k + 1$

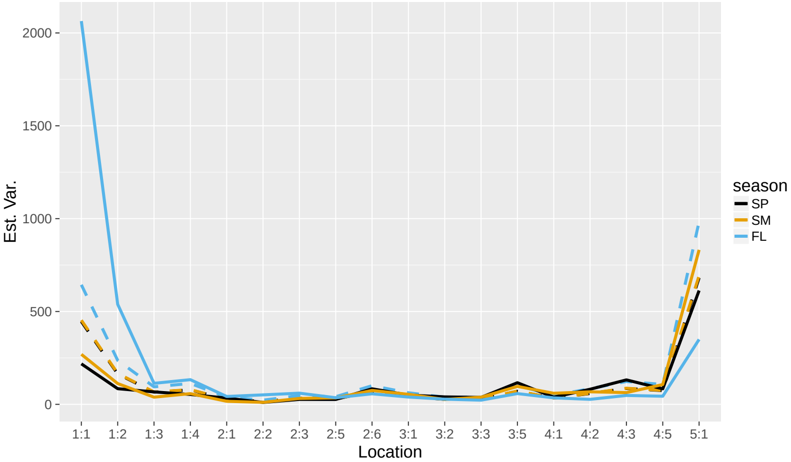
until $|\ell^k - \ell^{k-1}| \leq \epsilon$

Some (very) preliminary results

Estimates: Both methods the same?



Estimates: Both methods the same?



Some observations and things to do

- LRT rejects the Matrix Normal in our example ($p \approx 10^{-5}$)

Some observations and things to do

- LRT rejects the Matrix Normal in our example ($p \approx 10^{-5}$)
- Simulations indicate LRT has correct size for $n \approx 50$

Some observations and things to do

- LRT rejects the Matrix Normal in our example ($p \approx 10^{-5}$)
- Simulations indicate LRT has correct size for $n \approx 50$
- It is unknown when MLE of $\boldsymbol{\Sigma}$ exists and is unique ($n > r^2c^2 + p$ suffices)

Thank You!

References

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