Modeling multivariate data: from precipitation to finance

Oleg Makhnin
New Mexico Tech
Dept. of Mathematics

March 28, 2008

Abstract:

This talk will outline my ongoing and future planned projects, discussing possibilities for students' research. Two examples: spatial random field modeling (based on the MS work with Anandkumar Shetiya), and a future project with stochastic volatility modeling.

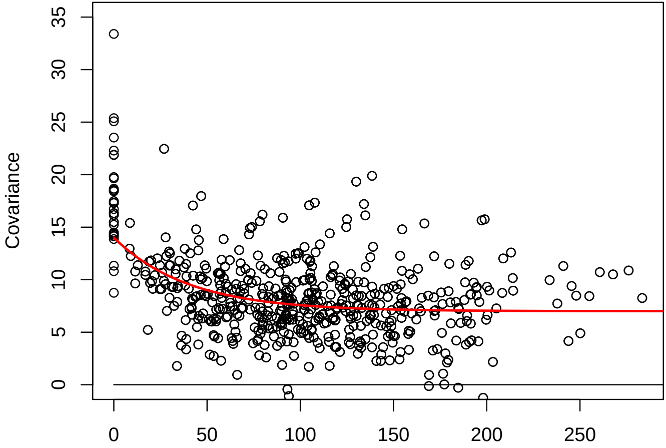
Intro

Multivariate problems

- Multiple variables are observed $Y^{(1)}, Y^{(2)}, \ldots$
- Each variable j is observed in time, that is $Y_1^{(j)}, Y_2^{(j)}, Y_3^{(j)}, \ldots$
- Have to deal with dependencies across time, as long as dependencies between the variables.
- Examples: precipitation measurements at gauge j, asset prices (asset $= j$, time $= t$ discretized)

Consider purely spatial dependency in precipitation.

Typically, if we plot covariance of precipitation measurements versus distance we observe something like this



Modeling spatial dependency

Need to model the dependency of $cov(Y^{(i)}, Y^{(j)})$ on the distance.

- • Traditional geostatistical methods based on variogram modeling: basically, model some function ϕ,

  $$cov(Y^{(i)}, Y^{(j)}) = \phi(\mathbf{r}_i - \mathbf{r}_j)$$

  $\mathbf{r}_j$ is the location of gauge j

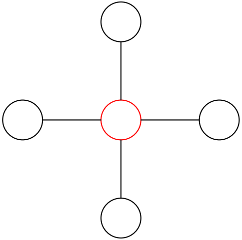
  (Math 586, Spring '09)
- • Convolution of GMRF's (Gaussian Markov random fields), which are easier to deal with computationally.

GMRF: Some basics

Markov field: (extension from 1-dim concept of Markov chain) possesses *Markov property*

$$P(X(\mathbf{s}_i)| \text{ rest of X}) = P(X(\mathbf{s}_i)|\{X(\mathbf{s}_j), j \sim i\})$$

i.e. the value at the location i depends only on the values at *neighboring* locations $j \sim i$. E.g. rectangular grid neighbors



GMRF: Some basics

Gaussian vector $\mathbf{x} = \{X(s_j), j = 1, ..., m\}$ is determined by its prior distribution

$$p(\mathbf{x}) \propto \exp(-\lambda_x \mathbf{x}^T V \mathbf{x}) \tag{1}$$

with *precision matrix* V defined by

$$V_{ij} = \begin{cases} -1, & i \sim j \\ \text{number of neighbors of } i, & i = j \\ 0, & \text{otherwise} \end{cases}$$

and λ_x is the smoothness parameter.
Advantage: matrix V is sparse.

GMRF: convolution

However, the gauges are not on a regular grid!

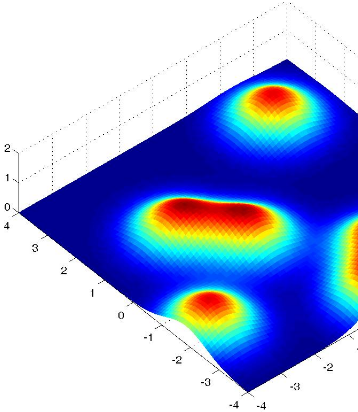
To fill in the values, use **convolutions**, at location $\mathbf{s}$

$$\tilde{X}(\mathbf{s}) = \sum_j X(\mathbf{s}_j)\psi(\mathbf{s} - \mathbf{s}_j)$$

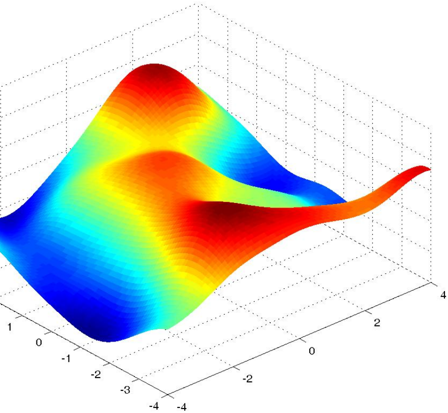
where ψ is some kernel function, say Gaussian

$$\psi(\mathbf{s}) = exp(-\mathbf{s}^T\mathbf{s}/2\sigma_\psi^2),$$

and σ_ψ can be taken equal to the grid spacing.



GMRF: convolution



Finally, our observations

$$Y(\mathbf{s}) = \sum_{j \in \texttt{grid}} X(\mathbf{s}_j)\psi(\mathbf{s}-\mathbf{s}_j)+e(\mathbf{s}) \tag{2}$$

where e are white-noise (i.i.d.) errors

GMRF: estimation

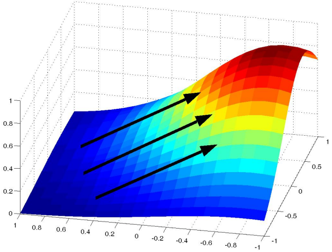
- Inverse problem: we need to estimate the unknown $\mathbf{x} = \{X(s_j)\}$, and other parameters $= \theta$, using the data.
- Handy tool for inverting: Bayesian approach.
  Given $p(\mathbf{x}, \theta)$= prior, and $p(\mathbf{y}|\mathbf{x}, \theta)$ = likelihood, our goal is to compute

  $$\textbf{Posterior:} \qquad p(\mathbf{x}, \theta|\mathbf{y}) \propto \texttt{prior} \cdot \texttt{likelihood}$$

  for the process X and other parameters of interest (like σ_e).
- Computational tools: Gibbs sampler.
  End up with a Markov Chain Monte Carlo (MCMC) sample.

MFD random field

Effects of terrain: MFD (moisture flux direction)
Increased precipitation when moisture is moving upslope.



Compute another random field W that helps in improving the prediction.

W is directional, i.e. its values are between 0 and 2π

MFD random field: prior

Try to follow the convolution approach.

Let vector $W =$ the values of MFD at the grid points.
We set a prior on W that encourages it to be smooth:

$$p(W) \propto \exp\left[\gamma \sum_{i \sim j} \cos(W_i - W_j)\right]$$

The values of W that are close (in circular topology) will receive a high prior weight, and diametrically opposite values of W receive a low prior weight.
$\gamma \geq 0$ is a smoothness parameter.

MFD random field: convolution

To find W co-located with the data Y, follow the convolution

$$\tilde{W}(\mathbf{s}) = \circ\sum_{j \in \texttt{grid}} \psi(\mathbf{s} - \mathbf{s}_j) W_j$$

however the sum taken above is understood in the circular topology, that is, through weighted averages

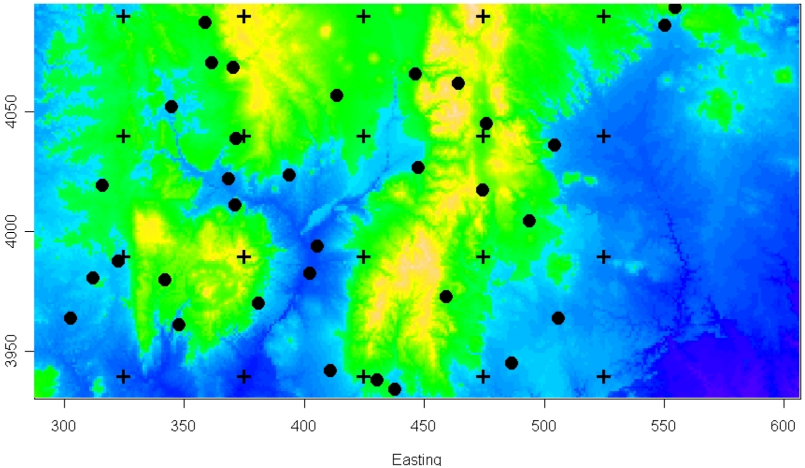
$$\sin(\tilde{W}(\mathbf{s})) = \frac{\sum_j \psi(\mathbf{s} - \mathbf{s}_j) \sin(W_j)}{\sum_j \psi(\mathbf{s} - \mathbf{s}_j)}, \qquad \cos(\tilde{W}(\mathbf{s})) = \ldots$$

The overall model becomes

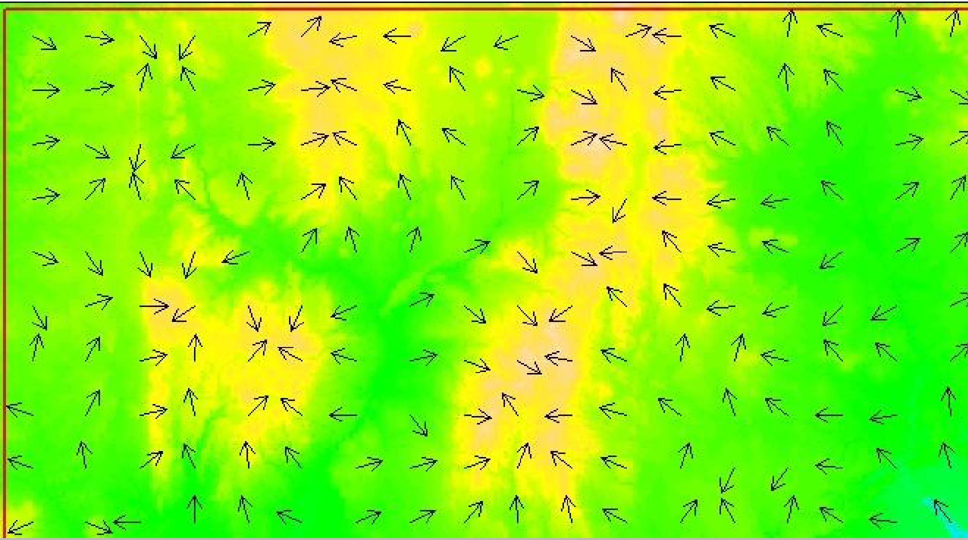
$$Y(\mathbf{s}) = \sum_j X(\mathbf{s}_j)\psi(\mathbf{s} - \mathbf{s}_j) + \beta\, cos(\tilde{W}(\mathbf{s}) - A(\mathbf{s})) + e(\mathbf{s}) \qquad (3)$$

$A(\mathbf{s})$ = terrain aspect at location $\mathbf{s}$

Study area

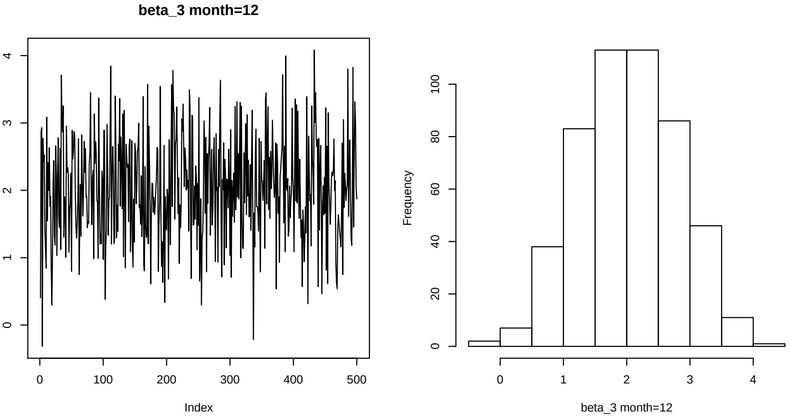


Study area: aspects

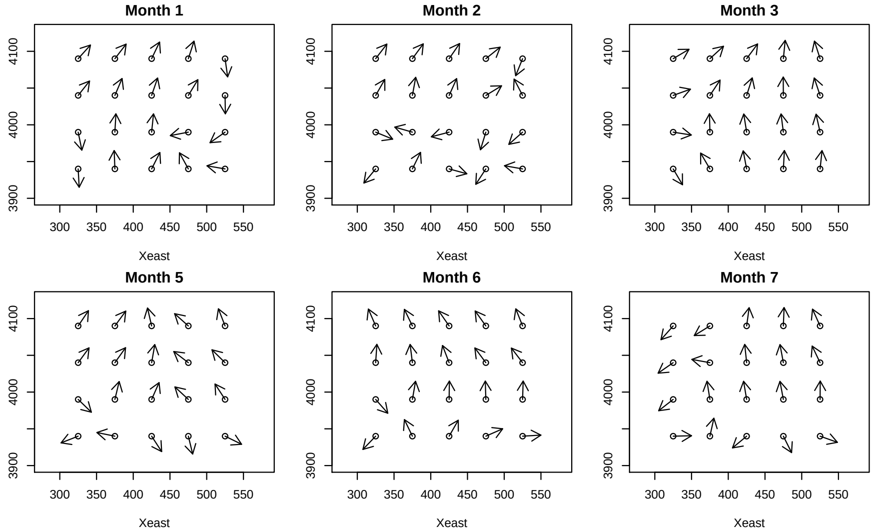


Output: β

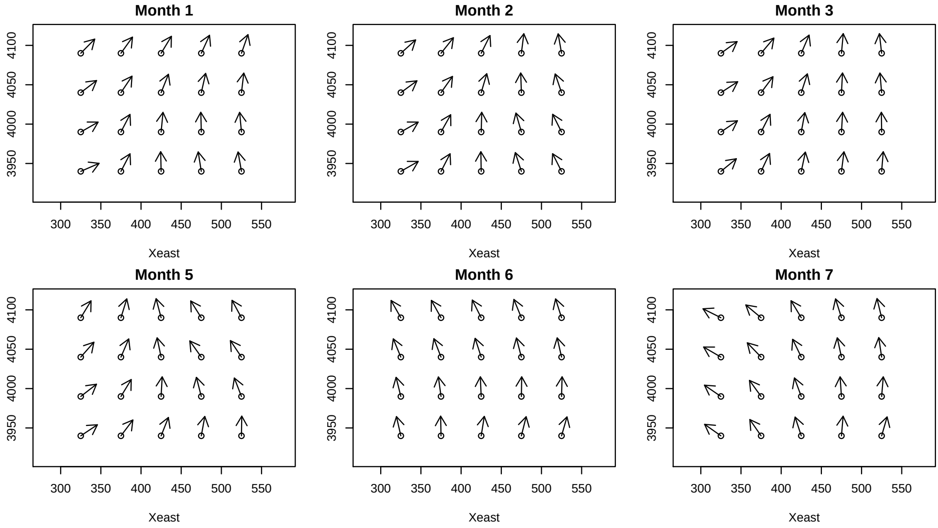
This shows a typical MCMC output for the posterior distribution of β.



Fitted MFD, $\gamma = 2$



Fitted MFD, $\gamma = 5$



Modeling financial data

Asset price model

Model the asset price X_t as a random walk

$$X_t = X_{t-1} + e_t$$

$\{e_t\}$ are independent

However, they are not identically distributed:

$$e_t \sim Normal(0, \sigma_t)$$

where $\sigma_t =$ "volatility" is itself time-varying; and it is predictable to some extent.

Stochastic volatility

Stochastic volatility model

This is an autoregressive model

$$log(\sigma_t) = \mu + r(log(\sigma_{t-1}) - \mu) + \nu_t$$

Research question: for different assets j, does the clustering of $\sigma_t^{(j)}$ take place?
That is, if we look at multiple assets, do they form the groups within which $\sigma_t^{(j)}$ behave alike?

Tools

- Multivariate Normal distribution
- Time-series approaches, autocorrelation, autoregression etc.
- Bayesian inference, computational methods (Markov Chain Monte Carlo)

QUESTIONS?

THANK YOU!

`www.nmt.edu/~olegm/talks/Multivar/`

for pdf file