Math 12230: Spatio-Temporal Statistics for Big Data Midterm Exam 2
Due: January 9, 2016 in the instructor's mailbox

1. (10 points) Consider ordinary spatio-temporal kriging where the process $Y(\cdot;\cdot)$ has constant unknown mean μ. Suppose we observe the data

$$Z(\mathbf{s}_i;t_{ij}) = Y(\mathbf{s}_i;t_{ij}) + \varepsilon(\mathbf{s}_i;t_{ij}), \quad j = 1,\dots,T_i,\ i = 1,\dots,m,$$

and need to predict $Y(\mathbf{s}_0;t_0)$. Let $\mathbf{Z}^{(i)} = (Z(\mathbf{s}_i;t_{i1}),\dots,Z(\mathbf{s}_i;t_{iT_i}))^T$, $\mathbf{Z} = (\mathbf{Z}^{(1)T},\dots,\mathbf{Z}^{(m)T})^T$, $\mathbf{C}_Z = \mathrm{Var}(\mathbf{Z})$, $\mathbf{c}_0 = \mathrm{Cov}(Y(\mathbf{s}_0;t_0),\mathbf{Z})$, and $C_{0,0} = \mathrm{Var}(Y(\mathbf{s}_0;t_0))$. Derive the ordinary-kriging predictor

$$\widehat{Y}(\mathbf{s}_0;t_0) = \{\mathbf{c}_0 + \mathbf{1}(1-\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{c}_0)/(\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{1})\}^T\mathbf{C}_Z^{-1}\mathbf{Z}$$

and variance

$$\sigma^2_{\mathrm{ok}}(\mathbf{s}_0;t_0) = C_{0,0} - \mathbf{c}_0^T\mathbf{C}_Z^{-1}\mathbf{c}_0 + (1-\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{c}_0)^2/(\mathbf{1}^T\mathbf{C}_Z^{-1}\mathbf{1}).$$

2. (10 points) Consider the linear dynamical spatio-temporal model

$$\mathbf{Z}_t = \mathbf{H}_t\mathbf{Y}_t + \boldsymbol{\varepsilon}_t,$$
$$\mathbf{Y}_t = \mathbf{M}_t\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are independent $N(\mathbf{0},\mathbf{R}_t)$ and $N(\mathbf{0},\mathbf{Q}_t)$, respectively. Derive the expressions of $\mathbf{Y}_{t|t-1}$, $\mathbf{P}_{t|t-1}$, $\mathbf{Y}_{t|t}$, and $\mathbf{P}_{t|t}$ in the forecast and filtering distributions

$$\mathbf{Y}_t|\mathbf{Z}_{1:t-1} \sim N(\mathbf{Y}_{t|t-1},\mathbf{P}_{t|t-1}),$$
$$\mathbf{Y}_t|\mathbf{Z}_{1:t} \sim N(\mathbf{Y}_{t|t},\mathbf{P}_{t|t}).$$

3. (20 points) Consider the S-species Lotka–Volterra model

$$\frac{dX_i}{dt} = X_i\left(r_i + \sum_{j=1}^{S} a_{ij}X_j\right), \quad i = 1,\dots,S,$$

where $\mathbf{r} = (r_1,\dots,r_S)^T$ and $\mathbf{A} = (a_{ij})$ are unknown parameters, and $a_{ii} < 0$.

   (a) Suppose we observe the data

$$Y_i(t_j) = X_i(t_j) + \varepsilon_i(t_j), \quad i = 1,\dots,S,\ j = 0,\dots,T,$$

   where $\varepsilon_i(t_j)$ are independent $N(0,\sigma^2)$. Describe a state-of-the-art method for estimating $\mathbf{r}$ and $\mathbf{A}$.

   (b) Suppose X_i are not observable; instead, we observe $\mathbf{Z}(t_j) \sim \mathrm{Multinomial}(n_j,\mathbf{p}_j)$, where $\mathbf{p}_j = (X_1(t_j),\dots,X_S(t_j))^T \big/ \sum_{i=1}^S X_i(t_j)$. Now are $\mathbf{r}$ and $\mathbf{A}$ identifiable? Can you suggest a method for estimating them? You are free to make assumptions such that the problem is tractable.

4. (20 points) Let $Z(\cdot)$ be a zero-mean Gaussian process with stationary, isotropic covariance function $C(\cdot;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is an unknown parameter. Consider the problem of estimating $\boldsymbol{\theta}$ from the data $\mathbf{Z} = (Z(\mathbf{s}_1),\dots,Z(\mathbf{s}_n))^T$ when n is very large.

   (a) Suggest a divide-and-conquer method for estimating $\boldsymbol{\theta}$.

   (b) Implement your method in R or any other programming language that you prefer, and test your code on simulated data. Provide all source code and necessary experimental details.

   (c) Comment on the computational and statistical efficiency of your method relative to the maximum likelihood estimator and its covariance-tapered version.