

A Modification of the EM Algorithm for Spatio-Temporal Models with Environmental Applications

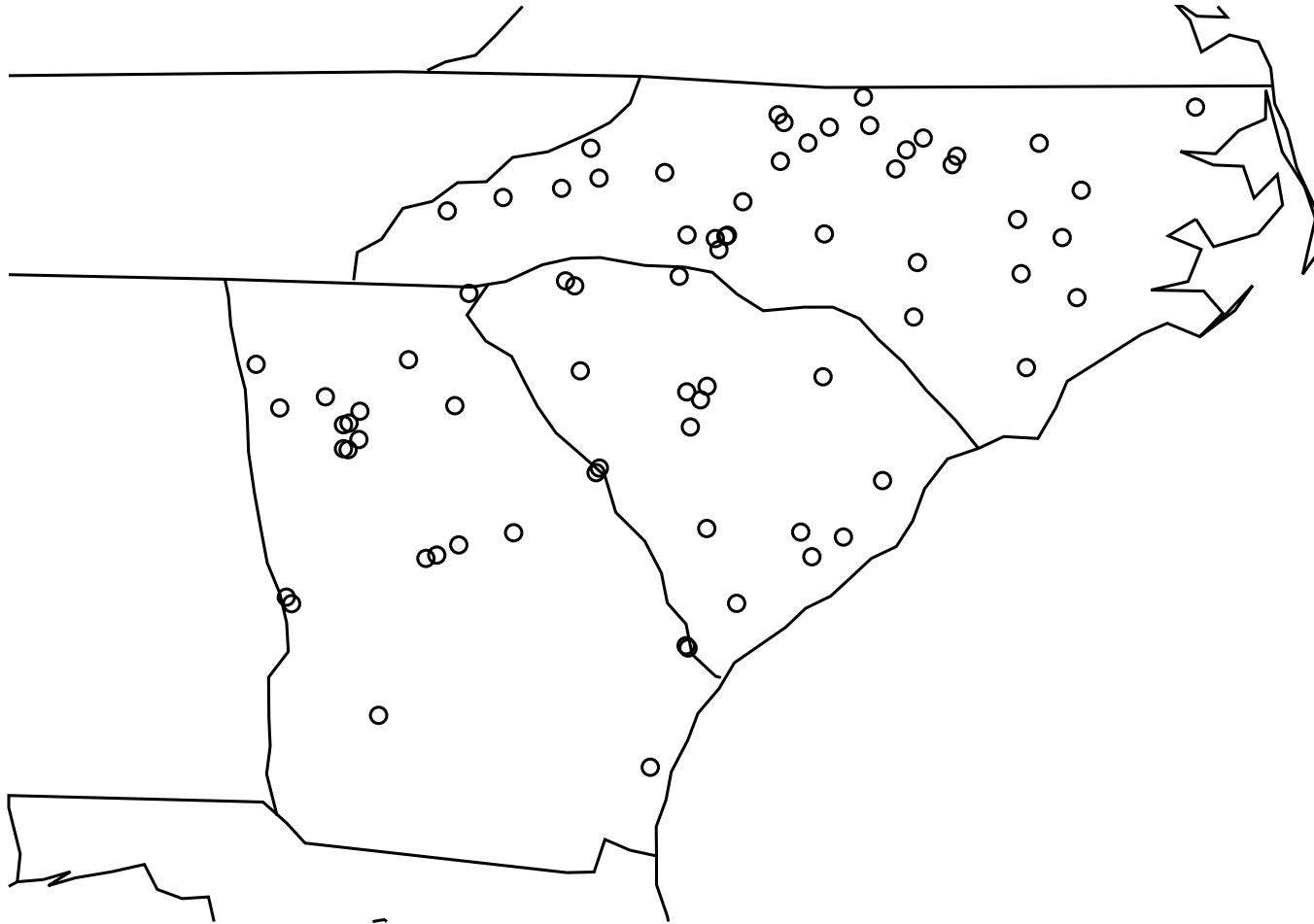
**Stanislav Kolenikov,
University of North Carolina, Chapel Hill
skolenik@unc.edu**

February 7, 2005

Structure of the talk

1. Motivation: EPA data on $PM_{2.5}$ (p. [3](#))
2. Background: spatial and spatio-temporal models (p. [6](#))
3. Background: EM algorithm (p. [15](#))
4. A modification of the EM algorithm and the implied likelihood (p. [34](#))
5. Application: results (p. [38](#))
6. Properties of the proposed approximation (p. [48](#), p. [54](#))

Motivation - I



EPA, 1999, PM_{2.5} monitors in NC, SC and GA

Motivation - II

Research questions and methods:

- Population exposure? EPA standard, long term average = $15 \mu\text{g}/\text{m}^3$
- Response variable transformation
- Generalized additive model: trends in space and time
- Spatial interpolation: kriging
- Parameter estimation: maximum likelihood, EM algorithm, a modification of the EM

Smith, Kolenikov & Cox (2003), described later (p. [38](#)).

Structure of the talk

1. Motivation: EPA data on $PM_{2.5}$ (p. [3](#))
2. Background: spatial and spatio-temporal models (p. [6](#))
3. Background: EM algorithm (p. [15](#))
4. A modification of the EM algorithm and the implied likelihood (p. [34](#))
5. Application: results (p. [38](#))
6. Properties of the proposed approximation (p. [48](#), p. [54](#))

Spatial statistics - I

Spatial statistics is concerned with the properties of the processes $Z(\mathbf{s})$ where typically $\mathbf{s} \in D \subset \mathbb{R}^2$.

- Given the measurements of ore concentration and deposition, what is the estimated deposition profile and total amount of ore in a field?
- Given the measurements of pollutant concentrations by a set of environmental monitors, what is the population exposure to that pollutant?
- Given a set of climatological measurements of different scales and modes (satellite/aircraft/land or ocean monitors), combine the data and use them as an input to a meteorological model
- Given the locations of the observed cases of an infectious disease, what is the epidemiological dynamics of the disease going to be?

Spatial statistics - II

More narrow area of *geostatistics* studies the properties of the covariance functions $\text{Cov}[Z(\mathbf{s}_k), Z(\mathbf{s}_l)]$ (Cressie 1993).

Convenient assumptions for a process $Z(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^d$:

$$\mathbb{E}[Z(\mathbf{s})] = \mu(\mathbf{s}), \quad (1)$$

$$\mathbb{V}[Z(\mathbf{s})] < \infty \quad (2)$$

Strict stationarity: $\forall \mathbf{h} \in \mathbb{R}^d$ and $\forall k, \mathbf{s}_1, \dots, \mathbf{s}_k \in D$ such that $\mathbf{s}_1 + \mathbf{h}, \dots, \mathbf{s}_k + \mathbf{h} \in D$, the distributions of the original and shifted data are the same:

$$Z(\mathbf{s}_1, \dots, \mathbf{s}_k) \stackrel{\mathcal{D}}{=} Z(\mathbf{s}_1 + \mathbf{h}, \dots, \mathbf{s}_k + \mathbf{h}) \quad (3)$$

Spatial statistics - III

Weak stationarity:

$$\forall \mathbf{s} \in D, \mu(\mathbf{s}) = \mu, \quad (4)$$

$$\forall \mathbf{s}_1, \mathbf{s}_2 \in D, \text{Cov}[Z(\mathbf{s}_1), Z(\mathbf{s}_2)] = C(\mathbf{s}_1 - \mathbf{s}_2) \quad (5)$$

for some function $C(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$.

Isotropy:

$$\forall \mathbf{s}_1, \mathbf{s}_2 \in D, \text{Cov}[Z(\mathbf{s}_1), Z(\mathbf{s}_2)] = C(\|\mathbf{s}_1 - \mathbf{s}_2\|) \quad (6)$$

for some function $C(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$.

The process said to be *Gaussian* if any finite sample from it has a multivariate normal distribution. For Gaussian processes, the two definitions of stationarity are equivalent.

Spatial statistics - IV

For a stationary process, define

$$\mathbb{V}[Z(\mathbf{s}_1) - Z(\mathbf{s}_2)] = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2) \quad (7)$$

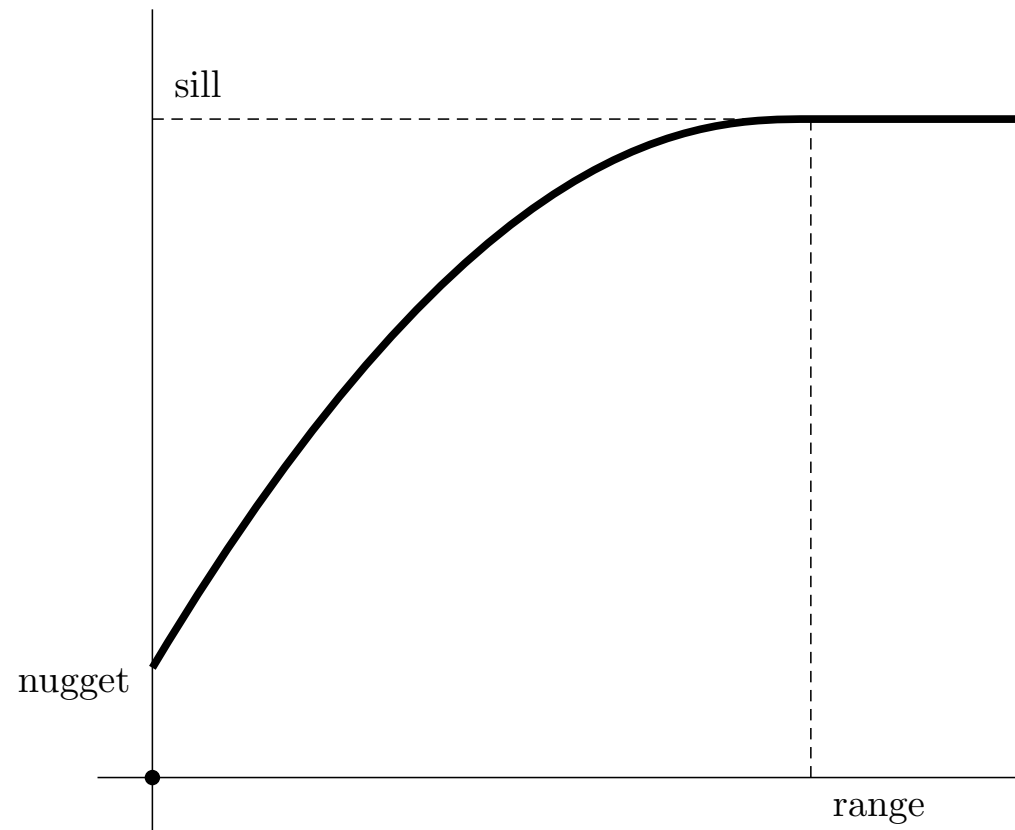
The function $\gamma(\cdot)$ is referred to as *semivariogram*, and $2\gamma(\cdot)$, as *variogram*.

Number of parametric forms available; see e.g. Smith (2003, Sec. 2.1).

[Click here for a nice table \(p. 12\).](#)

Intrinsic stationarity: the variogram representation of the covariance structure exists.

Spatial statistics - V



Typical shape of a variogram: *nugget*, *sill*, *range*

Spatial statistics - VI

Given a set of locations $\mathbf{s}_1, \dots, \mathbf{s}_n$ and observations of the spatial process $Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)$, what is the best linear unbiased predictor (BLUP) for a linear functional of the data?

$$W(Z) = \int_D w(\mathbf{s}) Z(\mathbf{s}) \nu(d\mathbf{s}) \quad (8)$$

where $w(\mathbf{s})$ is a kernel, and $\nu(\cdot)$ is a measure.

- $w(\mathbf{s}) = \delta(\mathbf{s} - \mathbf{s}_0) \Rightarrow W(Z) = Z(\mathbf{s}_0)$, a prediction for an unobserved location
- $w(\mathbf{s}) = 1/|D| \Rightarrow W(Z) =$ an areal average

Spatial statistics - VII

Universal kriging:

$$Z(\mathbf{s}_i) = X(\mathbf{s}_i)\beta + \eta(\mathbf{s}_i), \quad \eta \sim N(0, \Sigma), \quad i = 1, \dots, n,$$

$$\Sigma = \Sigma(\theta) \text{ is known,}$$

$$Z(\mathbf{s}_0) = x_0\beta + \eta_0 \text{ comes from the same field,} \quad \mathbb{E} \eta_0\eta = \tau \Rightarrow$$

$$\hat{z}_0 = x_0^T \hat{\beta} + \tau^T \Sigma^{-1} (Z - X^T \hat{\beta}), \quad (9)$$

$$\hat{\beta} \equiv \hat{\beta}_{GLS} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Z \quad (10)$$

$$\begin{aligned} \text{MSPE}[\hat{z}_0] &= \mathbb{V}[\hat{z}_0 - z_0] = \\ &= (x_0 - X^T \Sigma^{-1} \tau)^T (X^T \Sigma^{-1} X)^{-1} (x_0 - X^T \Sigma^{-1} \tau) - \\ &\quad - \tau^T \Sigma^{-1} \tau + \mathbb{V}[\eta_0] \end{aligned} \quad (11)$$

Spatial statistics - VIII

Components of (9) and (11)

- prediction from the linear regression part
- spatial correlation of the residuals
- unique error

Structure of the talk

1. Motivation: EPA data on $PM_{2.5}$ (p. [3](#))
2. Background: spatial and spatio-temporal models (p. [6](#))
3. Background: EM algorithm (p. [15](#))
4. A modification of the EM algorithm and the implied likelihood (p. [34](#))
5. Application: results (p. [38](#))
6. Properties of the proposed approximation (p. [48](#), p. [54](#))

Missing data - I

Are there any data sets that do not have missing data?..

The (complete) data generating mechanism:

$$X_i^c \sim \text{i.i.d. } f(\cdot; \theta), \quad i = 1, \dots, n \quad (12)$$

Some of the data are missing: $X^c = (X, X^m)$

The missing data:

$$Z_{ik} = \begin{cases} 1, & X_k \text{ is missing in } i\text{-th observation} \\ 0, & X_k \text{ is available in } i\text{-th observation} \end{cases} \quad (13)$$

The missing data mechanism is driven by parameters ψ .

Missing data - II

Typology (Little & Rubin 1987):

- X is *missing completely at random* (MCAR):

$$\Pr[Z_{ik} = 1|X, \psi] = p(\psi) \quad (14)$$

- X is *missing at random* (MAR):

$$\Pr[Z_{ik} = 1|X, \psi] = p(\psi, X_{-ik}) \quad (15)$$

- X is *not missing at random* (NMAR), or systematically missing:

$$\Pr[Z_{ik} = 1|X, \psi] = p(\psi, X_{-ik}, X_{ik}) \quad (16)$$

Missing data - III

The missing data mechanism is *ignorable*: the maximum likelihood estimation as if the data were complete still yields consistent and asymptotically efficient estimates if

1. the mechanism is MAR
2. θ and ψ are distinct.

Otherwise, one would need to build a model for the missing data, and use it in estimation.

EM algorithm - I

- A procedure to obtain the ML estimates when some of the data are missing
- An iterative maximization algorithm
- Bayesian flavor?
- Converges to the critical points of the likelihood surface
- Linear rate of convergence
- Implicitly assumes MAR

Dempster, Laird & Rubin (1977), McLachlan & Krishnan (1997)

EM algorithm - II

The EM stands for “expectation-maximization”

The expectation, or E-step: compute conditional expectation of the log-likelihood (or sufficient statistics) for the given observed data

$$Q(\theta; \theta^{(h)}) = \mathbb{E}_{\theta^{(h)}} [\ln L_c(\theta; X^c) | X, \theta^{(h)}] \quad (17)$$

The maximization, or M-step, is to maximize (17) w. r. t. θ :

$$\theta^{(h+1)} = \arg \max_{\theta} Q(\theta; \theta^{(h)}) \quad (18)$$

Iterate until convergence.

EM algorithm - III

1. On each complete iteration, the value of the likelihood is not decreased, hence
2. If the likelihood is bounded from above, the sequence of the EM iterations converges to a stationary point θ^* of the likelihood function.
3. The rate of convergence is linear:

$$\theta^{(h+1)} - \theta^* \approx J(\theta^*)(\theta^{(h)} - \theta^*) \quad (19)$$

where $J(\theta^*)$ is the Jacobian of the EM map, and hence the rate of convergence $0 < c < 1$ can be found through the spectral properties of the Jacobian, and depends on the proportion of missing data.

The linear rate of convergence is pretty slow, but the linearity can be used both to find the standard errors and to speed up the convergence.

EM algorithm - IV

GEM : generalized EM algorithms only attempt to somewhat increase the likelihood at each step rather than fully maximize it (that's enough to guarantee the convergence)

ECM : expectation and conditional maximization version splits the parameter space into several subspaces that span the whole space, and maximize separately over them

Monte Carlo : sample from the (conditional) distribution of the missing data and maximize over the parameters using (weighted) “replenished” data set

and others

EM algorithm - V

The primary problem statisticians would have with the EM algorithms is that it only produces point estimates without the standard errors.

- proportion of the missing data (very rough)?
- the bootstrap?
- supplemented EM algorithm: numerical derivatives of the likelihood function by only using the EM code.

Skip examples — go to p. [34](#).

EM Example 1

This example concerns the censored survival times. We have uncensored observations (y_1, \dots, y_r) and censored observations (y_{r+1}, \dots, y_n) (i.e. we know that in i -th experiment, $i > r$, the survival time is *at least* y_i). If the mean survival time is μ , then the density of y is

$$f(y; \mu) = \mu^{-1} \exp[-y/\mu], \quad y > 0 \quad (20)$$

and the survival function (probability of survival till time y)

$$S(y) = \exp[-y/\mu], \quad y > 0 \quad (21)$$

The likelihood is a combination of the two:

$$\ln L(\mu; \mathbf{y}) = - \sum_{i=1}^r \left(\ln \mu + \frac{y}{i\mu} \right) - \sum_{i=r+1}^n \frac{y}{i\mu = -r \ln \mu + \sum_{i=1}^n \frac{y}{i\mu}} \quad (22)$$

Then the ML estimate is $\hat{\mu}_{ML} = \sum y_i / r$.

EM Example 1 cont'd

Now, for the EM algorithm, we shall assume that the observations from $r + 1$ to n are observed. Then the “full” log-likelihood is

$\ln L = -n \ln \mu - \sum y_i / \mu$, and for the E-step, we replace the censored survival times with their best available estimates, $y_i + \mu^{(k+1)}$ (using the lack of memory property of the exponential distribution). Then

$$Q(\mu; \mu^{(k)}) = -n \ln \mu - \mu^{-1} \left[\sum_{i=1}^n y_i + (n - r) \mu^{(k)} \right] \quad (23)$$

and by maximizing it wrt μ on the M-step,

$$\mu^{(k+1)} = \frac{1}{n} \left[\sum_{i=1}^n y_i + (n - r) \mu^{(k)} \right] \quad (24)$$

The ML estimate is the unique solution of this equation,

$$\mu^{(k)} = \mu^{(k+1)} = \hat{\mu}_{ML}.$$

EM Example 2

In this example, we shall look at a bivariate normal distribution where some data are missing. Suppose we observe n observations from

$$y \sim N(\mu, \Sigma), \quad \mu = (\mu_1, \mu_2)^T, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

of which m are complete, m_1 have only y_1 observed, and m_2 have only y_2 observed.

EM Example 2 cont'd

The likelihood is complicated:

$$\begin{aligned}
 -2 \ln L(\theta; y) &\sim m \ln |\Sigma| + \sum_{i=1}^m (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \\
 &\quad + (m_1 \ln \sigma_{11} + m_2 \ln \sigma_{22}) + \\
 &\quad \left\{ \sum_{i=m+1}^{m+m_1} \frac{(y_1 - \mu_1)^2}{\sigma_{11}} + \sum_{i=m+m_1+1}^{m+m_1+m_2} \frac{(y_1 - \mu_1)^2}{\sigma_{22}} \right\} \quad (25)
 \end{aligned}$$

No trackable closed form solution.

EM Example 2 cont'd

With the complete data, one only needs the sufficient statistics

$$T_k = \sum_i y_{ik}, \quad T_{jk} = \sum_i y_{ij}y_{ik}, \quad j, k = 1, 2 \quad (26)$$

so that the parameter estimates are

$$\hat{\mu}_k = T_k/n, \quad \hat{\sigma}_{jk} = \frac{T_{jk} - n^{-1}T_jT_k}{n} \quad (27)$$

This is the essence of the M-step. At the E-step, we need to compute the expected values of those sufficient statistics based on the current parameter values and the observations.

EM Example 2 cont'd

If we are missing on y_2 , then given y_1 , the distribution of y_2 is $N(\mu_{2.1}, \sigma_{22.1})$ where

$$\mu_{2.1} = \mu_2 + \sigma_{12}\sigma_{11}^{-1}(y_1 - \mu_1), \quad \sigma_{22.1} = \sigma_{22}(1 - \rho^2) \quad (28)$$

Then the contributions to the sufficient statistics T_2, T_{22} will be $\mu_{2.1}$ and $\sigma_{22.1} + \mu_{2.1}^2$. Thus simply plugging $\mu_{2.1}$ into the likelihood (or equivalently imputing it into the data set) is insufficient: we also need the $\sigma_{22.1}$. This is an important nuance of the EM algorithm that distinguishes it from imputation procedures.

EM Example 3

In this example, we shall cluster the multivariate data, or, rather, estimate the parameters of the mixture of normal distributions.

The g -component normal mixture has a density

$$f(y, \theta) = \sum_{l=1}^g \pi_l f_l(y), \quad \sum_l \pi_l = 1, \quad f_l(y) = \frac{1}{\sigma_l} \phi\left(\frac{y - \mu_l}{\sigma_l}\right) \quad (29)$$

\implies (log) likelihood $L(\theta, \mathbf{y}) \implies \hat{\theta}_{ML} = (\hat{\pi}, \hat{\mu}, \hat{\sigma})$.

Here, no data is formally missing, so we shall introduce some. Namely, we shall consider the class labels as missing.

EM Example 3 cont'd.

Now, let us cast the problem in EM terms by adding an unobserved $z_{i,l} = \mathbb{I}\{i\text{-th observation belongs to } l\text{-th class}\}$. Then we have to deal with a multinomial likelihood:

$$\ln L(\theta; y) = \sum_{i,l} z_{i,l} \ln \pi_l + \sum_{i,l} z_{i,l} \ln \frac{1}{\sigma_l} \phi\left(\frac{y - \mu_l}{\sigma_l}\right) \quad (30)$$

On k -th iteration with the parameter estimates $\pi^{(k)}, \mu^{(k)}, \sigma^{(k)}$, the E-step is to compute the expected value of $Z_{i,l}$ given the data and $\theta^{(k)}$:

$$\mathbb{E}_{\theta^{(k)}} Z_{i,l} = \Pr\{Z_{i,l} = 1 | \theta^{(k)}, y_i\} = \pi^{(k)} f_l(y) / f(y) \equiv w_{i,l} \quad (31)$$

EM Example 3 continued.

The M-step is then to maximize the likelihood, and this is simply finding the weighted means and variances:

$$\mu_l^{(k+1)} = \frac{1}{n} \sum_{i=1}^n w_{i,l} y_i, \quad (32)$$

$$\sigma_l^{(k+1)} = \frac{1}{n} \sum_{i=1}^n w_{i,l} (y_i - \mu_l)^2 \quad (33)$$

$$\pi_l^{(k+1)} = \frac{1}{n} \sum_{i=1}^n w_{i,l} \quad (34)$$

The steps are repeated until “convergence”.

A clear advantage over k -means and similar methods is that the probabilities of classes are also obtained as the output of the procedure.

Structure of the talk

1. Motivation: EPA data on $PM_{2.5}$ (p. [3](#))
2. Background: spatial and spatio-temporal models (p. [6](#))
3. Background: EM algorithm (p. [15](#))
4. A modification of the EM algorithm and the implied likelihood (p. [34](#))
5. Application: results (p. [38](#))
6. Properties of the proposed approximation (p. [48](#), p. [54](#))

A modification of the EM algorithm - I

Suppose the complete data comes from a multivariate normal distribution:

$$Y_i^c \sim N(X_i^c \beta, \Sigma(\theta))$$

where β covers trends in space, time, and monitor characteristics, and $\Sigma(\theta)$ is the spatial covariance. Suppose some of the data are missing, so the observed data follow a multivariate normal distribution with the dimension d_i :

$$Y_i \sim N(X_i \beta, \Sigma_i)$$

Then

$$\begin{aligned} \ln L(\theta, \beta; \mathbf{y}, X) &\sim \\ &\sim -\frac{1}{2} \left\{ \sum_{i=1}^N \ln |\Sigma_i(\theta)| + \text{tr}[(\mathbf{y}_i - X_i \beta_i)(\mathbf{y}_i - X_i \beta_i)^T \Sigma_i(\theta)^{-1}] \right\} \quad (35) \end{aligned}$$

A modification of the EM algorithm - II

Sufficient statistic of the data: $y_i, y_i y_i^T, i = 1, \dots, N$. For the EM algorithm, we would need

$$\mathbb{E} \left[\sum_t (Y_i^c - X_i^c \beta)(Y_i^c - X_i^c \beta)^T | Y, X, \beta, \theta \right] \quad (36)$$

The exact EM: the expected value is found by kriging, see (9).

The approximate EM: use the marginal distributions $N(X_i \beta, \Sigma)$ instead of the conditionals given by (multivariate extensions of) (9)–(11). That is,

$$\begin{aligned} & \tilde{\mathbb{E}} \left[(Y_{ij}^c - X_{ij}^c \beta)(Y_{ik}^c - X_{ik}^c \beta) | Y, X, \beta, \theta \right] = \\ & = \begin{cases} (Y_{ij} - X_{ij} \beta)(Y_{ik} - X_{ik} \beta), & \text{both } j \text{ and } k \text{ are observed,} \\ \sigma_{jk}(\theta), & \text{otherwise} \end{cases} \quad (37) \end{aligned}$$

A modification of the EM algorithm - III

Approximate EM algorithm

- + Only one matrix, $\Sigma(\theta)$, need to be inverted per computation of the likelihood
- + Can use ECM to separate the trend from the covariance space
- Statistical properties unknown?

Structure of the talk

1. Motivation: EPA data on $PM_{2.5}$ (p. 3)
2. Background: spatial and spatio-temporal models (p. 6)
3. Background: EM algorithm (p. 15)
4. A modification of the EM algorithm and the implied likelihood (p. 34)
5. Application: results (p. 38)
6. Properties of the proposed approximation (p. 48, p. 54)

PM_{2.5} application - I

- US EPA 1999 data from some 780 continental US monitors;
- variables: PM_{2.5} concentration; latitude and longitude; the area type; altitude of the monitor; etc.
- frequency varies from daily to \approx weekly;
- lots of missing data

Restricted / collapsed data set: [74 monitors](#) (NC, SC, GA), 52 weeks — need for spatio-temporal modelling?

PM_{2.5} application - II

Spatial and temporal components are separated in the generalized additive model (Hastie & Tibshirani 1990) manner:

$$g(y_{it}) = \phi_{\text{space}}(i) + \phi_{\text{temp}}(t) + \phi_{\text{area}}(i) + \varepsilon_{it}, \quad (38)$$

- $g(\cdot)$ allows for the overall transformation to reduce skewness and/or to stabilize variance,
- $\phi_{\text{space}}(i)$ is the spatial trend (thin plate splines),
- $\phi_{\text{temp}}(t)$ is the temporal trend (B-splines),
- $\phi_{\text{area}}(i)$ is the additive term for the area type,
- the error terms ε_{it} are assumed to be uncorrelated over time, but correlated over space, so that $\forall t \text{ Cov } \varepsilon_t = \Omega$.

Skip to the results, p. [46](#).

PM_{2.5} application - III

The need for transformation can be assessed from the [variance vs. mean plots](#). The graph in levels clearly shows the increase in variance, while the square root transform graph shows a more stable variance.

Alternatively, we might have used the formal Box-Cox framework to obtain pretty much the same results.

Minor implication: the model additive in levels rather than in square roots or logs might be easier to understand as the model with sources, drains, and transport of the PM.

PM_{2.5} application - IV

The [trend in time](#) is modelled through B-splines

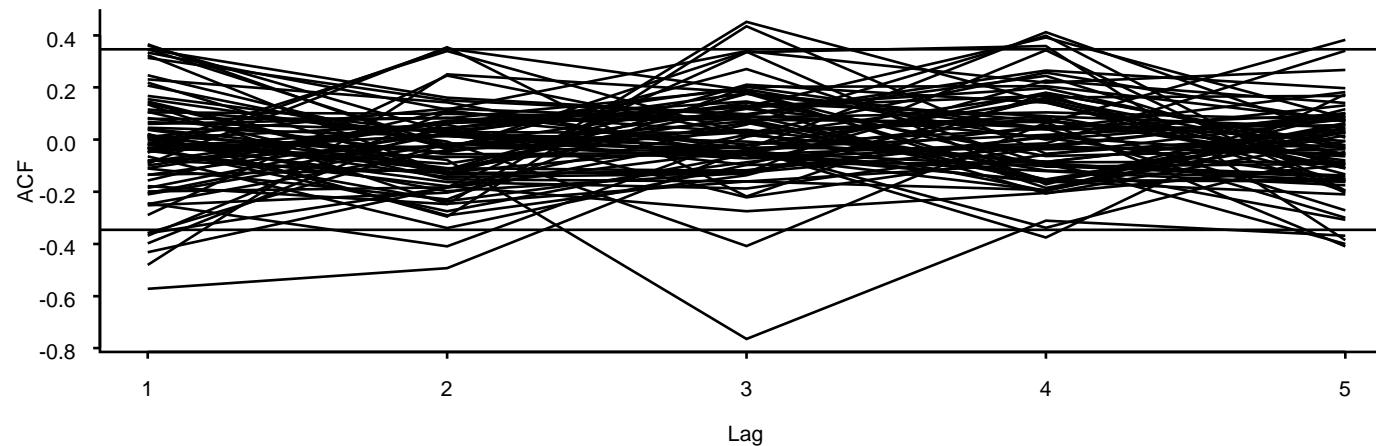
$$B(u) = \begin{cases} \frac{3|u|^3 - 6u^2 + 4}{6}, & -1 \leq u \leq 1, \\ \frac{(2 - |u|)^3}{6}, & 1 < |u| \leq 2, \\ 0, & 2 < |u|. \end{cases} \quad (39)$$

$$\phi_{temp}(t) = \alpha_0 + \sum_{k=1}^K \alpha_k \delta_k(t), \quad t \in [0, T], \quad \delta_k(t) = B\left(\frac{K}{T} \left(t - \frac{Tk}{K}\right)\right) \quad (40)$$

or by simple weekly averages (which was our final choice).

PM_{2.5} application - V

After we have modelled the time trend as weekly averages, the following autocorrelation plot seems to indicate we don't have substantial problems with time correlations.



Residual autocorrelation with 95% confidence bands.

PM_{2.5} application - VI

Spatial trend is approximated by the thin plate spline expansion:

$$\begin{aligned}\phi_{spatial}(\mathbf{z}) &= \beta_x x + \beta_y y + \sum_{j=1}^J \beta_j \psi(z_1 - x^{(j)}, z_2 - y^{(j)}) \\ \psi(x, y) &= \frac{r \log r}{16\pi}; \quad r = \sqrt{x^2 + y^2}\end{aligned}\tag{41}$$

of nodes J :

- full sample?
- a subsample of sites?
- clustering: a few nodes with $z^{(j)}$ somewhere in between

PM_{2.5} application - VII

A quick look at the [variograms](#) shows that the process is non-stationary, so the following *generalized* covariance model may be appropriate:

$$2\gamma(d(i, j)) = \mathbb{V}[\epsilon_{ti} - \epsilon_{tj}] = (1 - \delta_{ij})\alpha(\kappa + d(i, j))^p \quad (42)$$

$\delta_{ij} = \mathbb{I}\{i = j\}$, Kronecker's delta;

α : variance; > 0 ;

$d(\cdot, \cdot)$ distance between the sites;

p : the power (shape) parameter;

κ : the nugget effect

PM_{2.5} application - VIII

The parameter vector:

- spatial trend, thin plate spline basis expansion
- temporal trend, B-spline expansion
- land use of the site,
- variogram / spatial covariance matrix

PM_{2.5} application - IX

Estimates of the semivariogram parameters (see (42))

Method	κ	p	α
MLE			
Point estimate	2.06	0.92	0.061
Standard error	0.35	0.097	0.0017
Approximate EM			
Point estimate	2.13	0.92	0.049
Standard error	0.29	0.083	0.0012
Corrected s.e.	0.35	0.098	0.0019

See also some [nice maps of the PM_{2.5} concentrations](#).

Structure of the talk

1. Motivation: EPA data on $PM_{2.5}$ (p. [3](#))
2. Background: spatial and spatio-temporal models (p. [6](#))
3. Background: EM algorithm (p. [15](#))
4. A modification of the EM algorithm and the implied likelihood (p. [34](#))
5. Application: results (p. [38](#))
6. Properties of the proposed approximation (p. [48](#), p. [54](#))

Approximate EM: AR(1) process - I

A simple and analytically tractable example: Gaussian AR(1) process,

$$y_t = a + \rho y_{t-1} + \epsilon_t, \quad |\rho| < 1, \epsilon_t \sim \text{i.i.d. } N(0, \sigma_\epsilon^2) \quad (43)$$

Kriging predictor:

$$\mathbb{E}[y_t | y_{t-1}, y_{t+1}, \theta] = \mu + \frac{\rho}{1 + \rho^2} [(y_{t-1} - \mu) + (y_{t+1} - \mu)], \quad (44)$$

$$\mathbb{V}[y_t | y_{t-1}, y_{t+1}, \theta] = \frac{(1 - \rho^2)\sigma_\epsilon^2}{1 + \rho^2} \quad (45)$$

Approximate EM: AR(1) - II

Approximate EM:

$$\tilde{\mathbb{E}}[y_t | y_{t-1}, y_{t+1}, \theta] = \mu, \quad (46)$$

$$\tilde{\mathbb{V}}[y_t | y_{t-1}, y_{t+1}, \theta] = \frac{\sigma_\epsilon^2}{1 - \rho^2}, \quad (47)$$

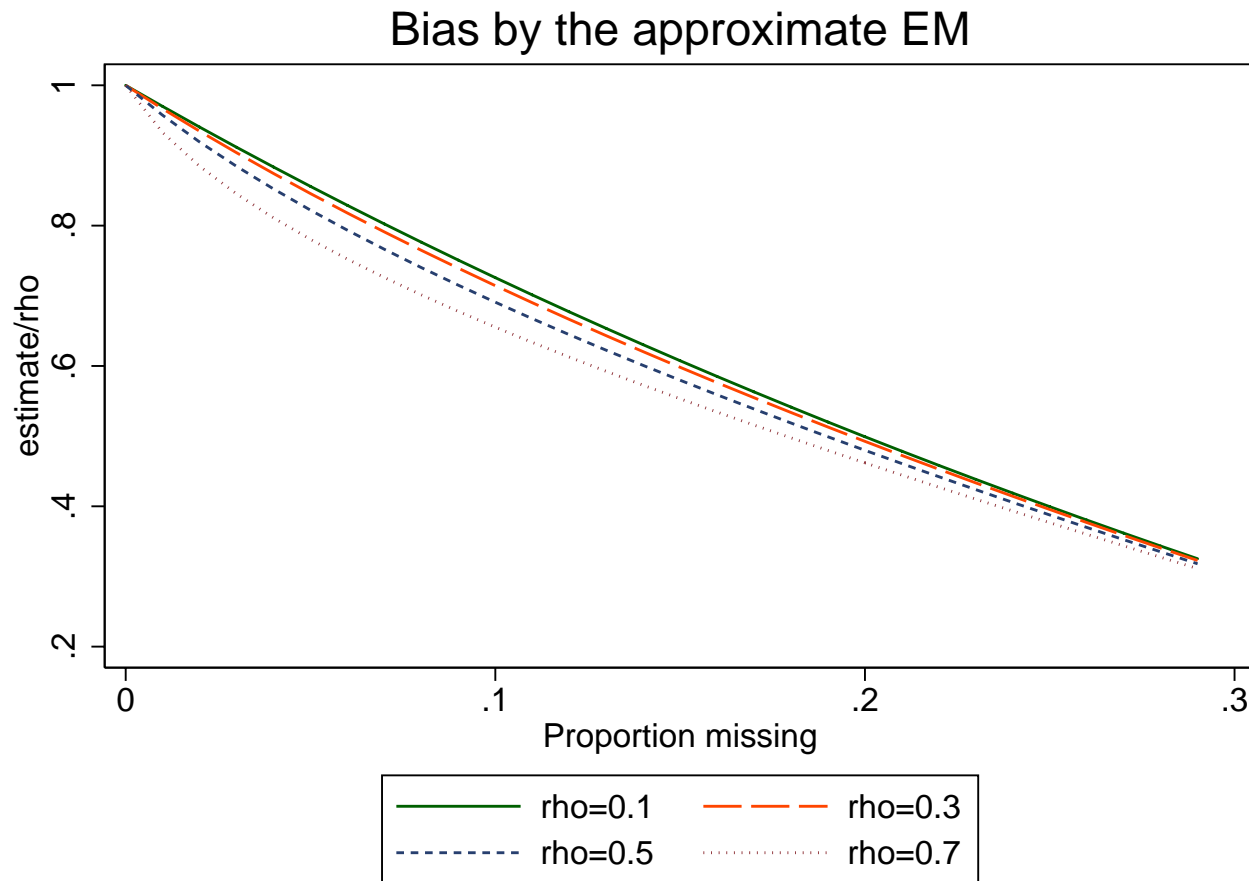
$$\tilde{\mathbb{E}}[(y_t - \mu)(y_{t-1} - \mu) | y_{t-1}, y_{t+1}, \theta] = \frac{\rho\sigma_\epsilon^2}{1 - \rho^2} \quad (48)$$

Approximate EM: AR(1) - III

Results:

- $\hat{\mu}$ is OK
- ρ is underestimated
- σ_ϵ^2 is underestimated from MLE, but consistently estimated from the sums of squares given a consistent estimate of ρ
- corrections for ρ are possible

Approximate EM: AR(1) - IV



Relative bias of $\hat{\rho}$

Approximate EM: AR(1) - V

Corrections to restore consistency of ρ :

- add a correction term

$$\mathcal{P}(\rho) = \nu T \left[\frac{1}{1 - \rho^2} - \frac{1 - \rho^2}{2} \right] \quad (49)$$

to the likelihood when it is maximized over ρ , where T is the length of the time series, and ν is the proportion of missing data,

or...

Approximate EM: AR(1) - V

- from the previous figure, the bias is approximately proportional to ν :

$$\text{plim } \hat{\rho} = \rho + A\nu + o(\nu),$$

$$A = -\frac{2 + (1 - \rho)^2}{(1 - \rho)^2} \rho,$$

$$\text{plim } \frac{\hat{\rho}}{\rho} = 1 - 3\nu + o(\nu) + o(\rho),$$

$$\tilde{\rho} = \hat{\rho}(1 - 3\nu)^{-1} \tag{50}$$

Approximate EM: general case - I

Complete data likelihood:

$$\begin{aligned}
 l(\theta, \beta; Y^c, X) &= -\frac{dN}{2} \ln 2\pi - \frac{N}{2} \ln |\Sigma(\theta)| - \\
 &\quad -\frac{1}{2} \sum_{i=1}^N \text{tr} \left[(Y_i^c - X_i^c \beta)(Y_i^c - X_i^c \beta)^T \Sigma^{-1} \right]
 \end{aligned} \tag{51}$$

Observed data likelihood:

$$\begin{aligned}
 l(\theta, \beta; Y, X) &= \\
 &= -\frac{1}{2} \sum_{i=1}^N \left\{ d_i \ln 2\pi + \ln |\Sigma_i(\theta)| + \text{tr} \left[(Y_i - X_i \beta)(Y_i - X_i \beta)^T \Sigma_i^{-1} \right] \right\}
 \end{aligned} \tag{52}$$

Approximate EM: general case - II

Presentness matrices P_i and *missingness* matrices M_i :

$$\begin{aligned} Y_i &= P_i Y_i^c, & Y_i^m &= M_i Y_i^c \\ Y_i^c &= P_i^T Y_i + M_i^T Y_i^m, \\ \Sigma_i &= P_i \Sigma(\theta) P_i^T \end{aligned} \tag{53}$$

Approximate EM: general case - III

The likelihood of the full data set, under the assumption of independence over i , is

$$\begin{aligned}
 l(\theta, \beta; Y, X) = & -\frac{1}{2} \ln 2\pi \sum_{i=1}^N d_{i,o} - \frac{1}{2} \sum_{i=1}^N \ln |P_i \Sigma(\theta) P_i^T| - \\
 & - \frac{1}{2} \sum_{i=1}^N \text{tr} \left[(P_i \Sigma P_i^T)^{-1} (Y_i^c - P_i X_i \beta) (Y_i^c - P_i X_i \beta)^T \right] \quad (54)
 \end{aligned}$$

Approximate EM/likelihood:

$$\begin{aligned}
 \tilde{l}(\theta, \beta; Y, X) = & -\frac{dN}{2} (\ln 2\pi + 1) - \frac{N}{2} \ln |\Sigma(\theta)| - \\
 & - \frac{1}{2} \sum_{i=1}^N \text{tr} \left\{ P_i \Sigma^{-1} P_i^T \left[(Y_i^o - P_i X_i \beta) (Y_i^o - P_i X_i \beta)^T - P_i \Sigma P_i^T \right] \right\} \quad (55)
 \end{aligned}$$

Approximate EM: general case - IV

Aim: derive estimating equations

- for the maximum likelihood estimates
- for the approximate EM

and compare the properties of the resulting estimates

Background: matrix calculus - I

Differential of a matrix function: the linear part of the increment

Let $F : S \rightarrow \mathbb{R}^{m \times p}$ be a matrix function defined on $S \subset \mathbb{R}^{n \times q}$. Let matrix $C \in \text{int } S$, and let $U \in \mathbb{R}^{n \times q}$ be such that $\|U\| < r$ (i.e., $U \in B(0, r)$, an open ball of radius r centered at zero with respect to the spectral norm), so that $C + U \in B(C, r) \in S$. If there exists a real matrix A of size $mp \times nq$ that depends on C , but not on U , such that

$$\text{vec}[F(C + U)] = \text{vec}[F(C)] + A(C) \text{vec}[U] + \text{vec}[R_C(U)] \quad \forall U \in B(C, r) \quad (56)$$

and

$$\lim_{U \rightarrow 0} \frac{R_C(U)}{\|U\|} = 0 \quad (57)$$

then F is said to be *differentiable at C* , and $m \times p$ matrix $\mathbf{d}F(C; U)$ given by

$$\text{vec}[\mathbf{d}F(C; U)] = A(C) \text{vec}[U] \quad (58)$$

is the *first differential of F at C with an increment U* , and $mp \times nq$ matrix $A(C)$ is the *(first) derivative of F at C* .

Background: matrix calculus - II

$$d(U + V) = dU + dV,$$

$$d(\alpha U) = \alpha dU,$$

$$dU^T = (dU)^T,$$

$$d \operatorname{vec}[U] = \operatorname{vec}[dU],$$

$$d(UV) = (dU)V + U(dV),$$

$$d \operatorname{tr} U = \operatorname{tr} dU$$

$$d(U \otimes V) = (dU) \otimes V + U \otimes (dV)$$

If additionally U is a non-degenerate square matrix, $|U| \neq 0$, then

$$d|U| = |U| \operatorname{tr}[U^{-1} dU],$$

$$d \ln |U| = \operatorname{tr}[U^{-1} dU],$$

$$dU^{-1} = -U^{-1}(dU)U^{-1} \tag{59}$$

Magnus & Neudecker (1999)

MLE: general case - V

Maximum likelihood estimating equations:

$$\begin{aligned}
 & d l(\theta, \beta; Y, X) = \\
 & = \frac{1}{2} \sum_{i=1}^N \text{tr} \left\{ (P_i \Sigma P_i^T)^{-1} \left[P_i \{d \Sigma\} P_i^T (P_i \Sigma P_i^T)^{-1} R_i - \right. \right. \\
 & \quad \left. \left. - 2 P_i X_i \{d \beta\} (Y_i - P_i X_i \beta)^T \right] \right\} \tag{60}
 \end{aligned}$$

where R_i is the matrix residual:

$$R_i = (Y_i - P_i X_i \beta)(Y_i - P_i X_i \beta)^T - P_i \Sigma P_i^T \tag{61}$$

MLE: general case - VI

Regression parameter subspace:

$$\begin{aligned}
 0 &= \sum_{i=1}^N \text{tr} \left[(P_i \Sigma P_i^T)^{-1} P_i X_i \{d\beta\} (Y_i - P_i X_i \beta)^T \right] = \\
 &= \sum_{i=1}^N \text{tr} \left[(Y_i - P_i X_i \beta)^T (P_i \Sigma P_i^T)^{-1} P_i X_i \{d\beta\} \right] = \\
 &= \sum_{i=1}^N \left[\{d\beta\}^T X_i^T P_i^T (P_i \Sigma P_i^T)^{-1} (Y_i - P_i X_i \beta) \right], \quad (62)
 \end{aligned}$$

$$\hat{\beta} \equiv \hat{\beta}_{\text{GLS}} = \left\{ \sum_{i=1}^N X_i^T P_i^T (P_i \Sigma P_i^T)^{-1} P_i X_i \right\}^{-1} \left\{ \sum_{i=1}^N X_i^T P_i^T (P_i \Sigma P_i^T)^{-1} Y_i \right\} \quad (63)$$

MLE: general case - VII

The estimating equation for the covariance parameters will be more complex. If

$$\Sigma(\alpha, \kappa, \psi) = \alpha C(\kappa, \psi), \quad (64)$$

κ is the nugget effect, and (possibly a vector) ψ describes the spatial correlation, then

$$d \Sigma(\theta) = d\alpha C(\kappa, \psi) + \alpha d\kappa I_d + \sum_j \alpha C_j(\psi) d\psi_j \quad (65)$$

where $C_j(\psi)$ is the matrix with zero on diagonal, and k, l -th off-diagonal entry equal to $\frac{\partial \rho(\psi, k, l)}{\partial \psi_j}$.

MLE: general case - VIII

Then the estimating equations for the covariance parameter subspace are

$$\frac{\partial l(\theta, \beta; Y, X)}{\partial \alpha} = \frac{1}{2} \sum_{i=1}^N \text{tr} \left[(P_i \Sigma P_i^T)^{-1} P_i C(\kappa, \psi) P_i^T (P_i \Sigma P_i^T)^{-1} R_i \right], \quad (66)$$

$$\frac{\partial l(\theta, \beta; Y, X)}{\partial \kappa} = \frac{1}{2} \sum_{i=1}^N \text{tr} \left[(P_i \Sigma P_i^T)^{-1} P_i \alpha P_i^T (P_i \Sigma P_i^T)^{-1} R_i \right], \quad (67)$$

$$\frac{\partial l(\theta, \beta; Y, X)}{\partial \psi_j} = \frac{1}{2} \sum_{i=1}^N \text{tr} \left[(P_i \Sigma P_i^T)^{-1} P_i \alpha C_j(\psi) P_i^T (P_i \Sigma P_i^T)^{-1} R_i \right] \quad (68)$$

Note that $\hat{\beta} \perp (\hat{\alpha}, \hat{\kappa}, \hat{\rho})$.

Approximate EM: general case - IX

The estimating equations for the approximate likelihood:

$$d\tilde{l}(\theta, \beta; Y, X) = \frac{1}{2} \sum_{i=1}^N \left\{ -\text{tr}(\Sigma^{-1}\{d\Sigma\}) + \text{tr}(P_i\Sigma^{-1}\{d\Sigma\}\Sigma^{-1}P_i^T R_i + \right. \\ \left. + P_i\Sigma^{-1}P_i^T [2P_iX_i\{d\beta\}(Y_i - P_iX_i\beta)^T + P_i\{d\Sigma\}P_i^T]) \right\} \quad (69)$$

where R_i is the matrix residual:

$$R_i = (Y_i - P_iX_i\beta)(Y_i - P_iX_i\beta)^T - P_i\Sigma P_i^T$$

Bias terms:

$$B(\Sigma, d\Sigma) = \sum_{i=1}^N [\text{tr}(P_i\Sigma^{-1}P_i^T P_i\{d\Sigma\}P_i^T) - \text{tr}(\Sigma^{-1}\{d\Sigma\})] \quad (70)$$

Sampling from a matrix

The matrices P_i sample rows and columns of the matrices they are applied to. If the proportion of the missing data is ν , then the diagonal entries of the corresponding covariance matrix of the data are sampled at the rate $1 - \nu$, and the off-diagonal entries, at a rate $(1 - \nu)^2$. If \mathbb{E}_s is the expectation with respect to the missing data mechanism (i.e., sampling of the rows and columns), then

$$\begin{aligned}
 \mathbb{E}_s \operatorname{tr}(PAP^T PB^T P^T) &= \mathbb{E}_s \sum_{k \in s} \sum_{j \in s} a_{jk} b_{jk} = \\
 &= (1 - \nu)^2 \sum_{k=1}^d \sum_{j=1}^d a_{jk} b_{jk} + \nu(1 - \nu) \sum_{k=1}^d a_{kk} b_{kk} = \\
 &= (1 - \nu) \operatorname{tr}\{A[(1 - \nu)B^T + \nu \operatorname{diag} B]\} = \\
 &= (1 - \nu) \operatorname{tr}\{A[B^T - \nu(B^T - \operatorname{diag} B)]\} \tag{71}
 \end{aligned}$$

Approximate EM: general case - X

If the missing data process is MCAR, then for $N \rightarrow \infty$, by the law of large numbers,

$$\begin{aligned}
 & \frac{1}{N} \sum_{i=1}^N \text{tr}(P_i \Sigma^{-1} P_i^T P_i \{d \Sigma\} P_i^T) \xrightarrow{p} \\
 & (1 - \nu) \text{tr}\left\{ \Sigma^{-1} \left[(1 - \nu) d \Sigma + \nu \text{diag } d \Sigma \right] \right\} = \\
 & = (1 - \nu) \text{tr}\left\{ \Sigma^{-1} \left[\alpha d \kappa I_d + \nu d \alpha \text{diag } C(\kappa, \psi) + \right. \right. \\
 & \left. \left. + (1 - \nu) (d \alpha C(\kappa, \psi) + \sum_j \alpha C_j(\psi) d \psi_j) \right] \right\} \quad (72)
 \end{aligned}$$

where the probability limit is taken over repeated sampling of sites, assuming samples are independent for different i 's (which is a part of MCAR assumption).

Approximate EM: general case - XI

How can this result be used? Consider the estimating equation for κ :

$$\begin{aligned} \frac{\partial \tilde{l}}{\partial \kappa} &= \frac{\alpha}{2} \sum_{i=1}^N \left\{ -\text{tr}(\Sigma^{-1}) + \text{tr}(P_i \Sigma^{-1} P_i^T + P_i \Sigma^{-1} \Sigma^{-1} P_i^T R_i) \right\} \\ \xrightarrow{p} \frac{\alpha}{2} \sum_{i=1}^N \left\{ -\text{tr}(\Sigma^{-1}) + \text{tr}((1 - \nu) \Sigma^{-1}) + \text{tr}(P_i \Sigma^{-1} \Sigma^{-1} P_i^T R_i) \right\} \end{aligned} \quad (73)$$

If the term $\text{tr}(\Sigma^{-1})$ is attenuated by a factor of $1 - \nu$, then it will cancel out with the next term, and the estimating equation will become consistent. As long as it is a full derivative, the correction is to multiply $\ln |\Sigma(\theta)|$ by $1 - \nu$ when the likelihood is maximized over κ .

Approximate EM: general case - XII

Equation (72) implies different corrections for different parameters, as by (65), the components of $d\Sigma$ related to the scale, nugget and correlation structure will have different structure w.r.t. the diagonal. As long as the term $\text{tr}(\Sigma^{-1} d\Sigma)$ comes as a differential of $\ln |\Sigma(\theta)|$, the following corrections can be entertained:

- nugget effect κ : diagonal only; need to attenuate $\ln |\Sigma(\theta)|$ by $1 - \nu$
- correlation structure ρ , shape and range parameters: off-diagonal elements only, need to attenuate $\ln |\Sigma(\theta)|$ by $(1 - \nu)^2$
- scale α : the correction depends on other parameters; the estimate of α should be derived from the generalized sum of squares rather than from the nonlinear maximization.

Further work

- Simulation study
- Separable processes:

$$\text{Cov}[Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)] = C_s(\mathbf{s}_1, \mathbf{s}_2)C_t(t_1, t_2)$$

- “Design” of unbiased estimating equations — e.g., versions of least squares
- Prediction of the second moments from the principal components/EOF
- Using small neighborhood for (approximate) kriging

References

Cressie, N. (1993), *Statistics for Spatial Data*, 2nd edn, Wiley, New York.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), ‘Maximum likelihood from incomplete data via the EM algorithm (with discussion)’, *Journal of the Royal Statistical Society B* **39**, 1–38.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman & Hall/CRC.

Little, R. J. A. & Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, Wiley, New York.

Magnus, J. R. & Neudecker, H. (1999), *Matrix differential calculus with applications in statistics and econometrics*, 2nd edn, John Wiley & Sons.

McLachlan, G. G. & Krishnan, T. (1997), *The EM Algorithm and Extensions*, John Wiley and Sons, New York.

Smith, R., Kolenikov, S. & Cox, L. H. (2003), ‘Spatio-temporal modeling of PM_{2.5} data with missing values’, *Journal of Geophysical Research – Atmospheres* **128**(D24). 9004, doi:10.1029/2002JD002914.

Smith, R. L. (2003), Environmental statistics. Unpublished manuscript under revision for publication as a book.
<http://www.stat.unc.edu/postscript/rs/envnotes.ps>.