Part 1: Geostatistical processes

Last week:

- How to model spatial continuous processes.
  - Hierarchical models
  - Gaussian Random Field (GRF)
  - Covariance functions (and variograms)

This week:

- Predictors for $Y|Z$ (assume Z known) (Tuesday)
- Estimation of covariance function (parameters θ) / variogram.
- Simulations + more models.

Today:

- Revisit our daily precipitation modeling challenge.
- Evaluation
- Simulations / computational issues
- Empirical variogram
- Understanding models and splines

Hierachical statistical models (HM) for Mountain temperature

Temperature:

Data model: $[Z|Y,\theta]$ Observations given true temperature:
Independent $Z(s_i) = N(Y(s_i), \sigma_\epsilon^2)$

Process model: $[Y|\theta]$ Distribution for temperature.
$Y(s) = \beta_0 + \beta_1 h(s) + \delta(\theta)$ where $\delta \sim GRF$ with
$E(\delta) = 0$ and covariance function $C_Y(s_1, s_2)$

Parameter model: $[\theta]$ Prior for parameters. $[\theta] = [\sigma_\epsilon^2][\beta][C_Y()]$.

- $h(s)$ is elevation (meters above see level)
- Can write as matrices: $Y(s) = X\beta$ with $X = [1, h(s)]^T$ and $\beta = [\beta_0, \beta_1]^T$

Modeling challenge: Daily precipitation

SeNorge

1. Suggest a spatial HM for daily precipitation?
2. What model/method/covariates do you think is used at SeNorge? Why?
3. How would you fit your model (make inference)?
4. How would you evaluate your model?

What is used for interpolation for precipitation at SeNorge?

New Routines for Gridding of Temperature and Precipitation Observations for "seNorge.no"

For the spatial interpolation of precipitation, the method of triangulation is used. Moreover, gridded precipitation values are corrected for the altitude of the respective seNorge grid point, using a vertical precipitation gradient of 10% *per 100 m height difference below an altitude of 1000 m above sea level as well as a gradient of* 5% *per 100 m height difference above an altitude of 1000 m above sea level (Tveito et al., 2000).*

Evaluation / diagnostics (Sec 2.2.2)

- Does our HM fit the data?
- How good are the predictions?

Model fit:
- AIC, BIC, DIC, Bayes factors
- Posterior predictive p-values

Predictions:
- Validation (test and training set)
- Cross-validation (use the dataset many times into training and test sets)

Do you see any challenges with the SeNorge case using cross-validation?

Inference about θ and Y. Computational cost?

What are the expensive parts of evaluating MVN and finding conditional MVN? (see MVN slide)

- $|\Sigma|$
- Σ^{-1}

Computational cost? $O(m^3)$ with m observations

- Much research on how to decrease computational burden.
- One appraoch is to work with sparse matrices (SPDE-appraoch, or lattice models in part 2)

Multivariate Normal distribution

Multivariate Normal(MVN) density

$Y = (Y_1, Y_2, \ldots, Y_n)$ is MVN with expected value μ and covarance Σ, $Y \sim MVN(\mu, \Sigma)$ if

$$f(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(t-\mu))$$

Conditional MVN

Let $Y = (Y_1, Y_2)^T$, $\mu = (\mu_1, \mu_2)^T$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then the $[Y_1|Y_2 = a] \sim MVN(\mu_{1|2}, \Sigma_{1|2})$ with

- $\mu_{1|2} = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$ and
- $\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Example: Temperature difference (pg 134-135)

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \delta(\mathbf{s}), \qquad \mathbf{s} \in D_s,$$

where for $\mathbf{s} = (s_1, s_2)'$, $\mathbf{x}(\mathbf{s}) = (1, s_1, s_2, s_2^2, s_2^3)'$, and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$

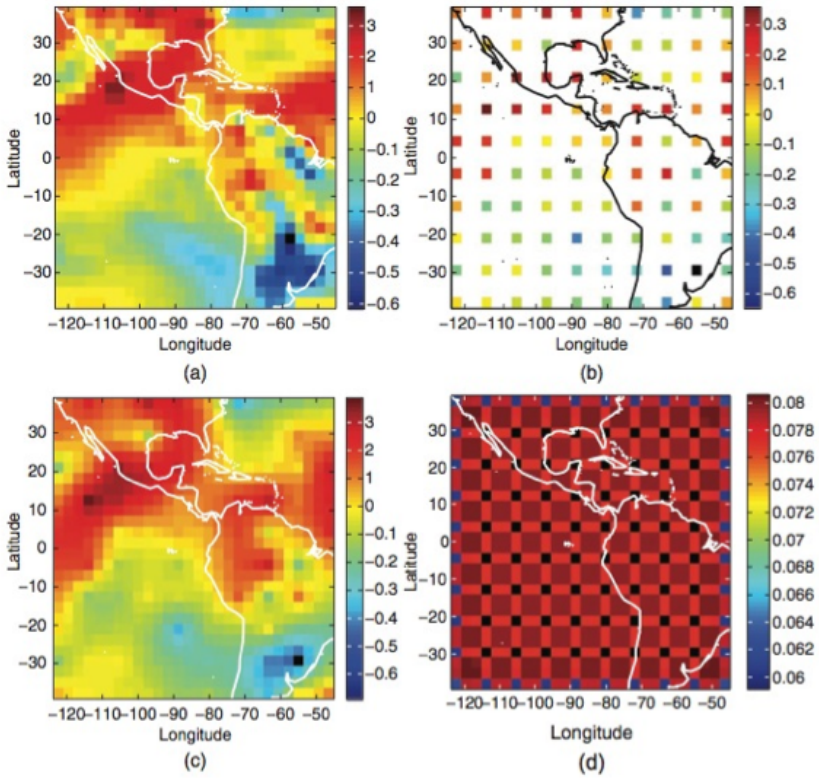

Figure 4.6 **(a)** Same plot as Figure 4.5a: Temperature change (1990s minus 1980s); the 28×28 values represent "the truth." **(b)** The 10×10 observations are obtained by subsampling "the truth" and adding mean-zero Gaussian noise. **(c)** Simple kriging predictor obtained from the missing and noisy data in **(b)**. **(d)** Kriging standard error corresponding to simple kriging; the pattern is expected due to regular sampling in **(b)**.

Variogram and empirical variogram

- An alternative to covariance functions measuring spatial dependency.

Stationary variogram

$$\gamma_Y(h) = \frac{1}{2} Var(Y(s+h) - Y(s))$$

computed from the data:

$$2\hat{\gamma}_Z^o(h) \equiv \text{ave}\{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : \|\mathbf{s}_i - \mathbf{s}_j\| \in T(h); i, j = 1, \ldots, m\}, \quad (4.16)$$

where $T(h)$ is a tolerance region around h (such as $h \pm \Delta$, Δ small). The *empirical semivariogram* is $\hat{\gamma}_Z^o(\cdot)$. For (4.16) to be an *appropriate estimator*

- Is an estimate of the variogram of Z.

Example emprical variogram

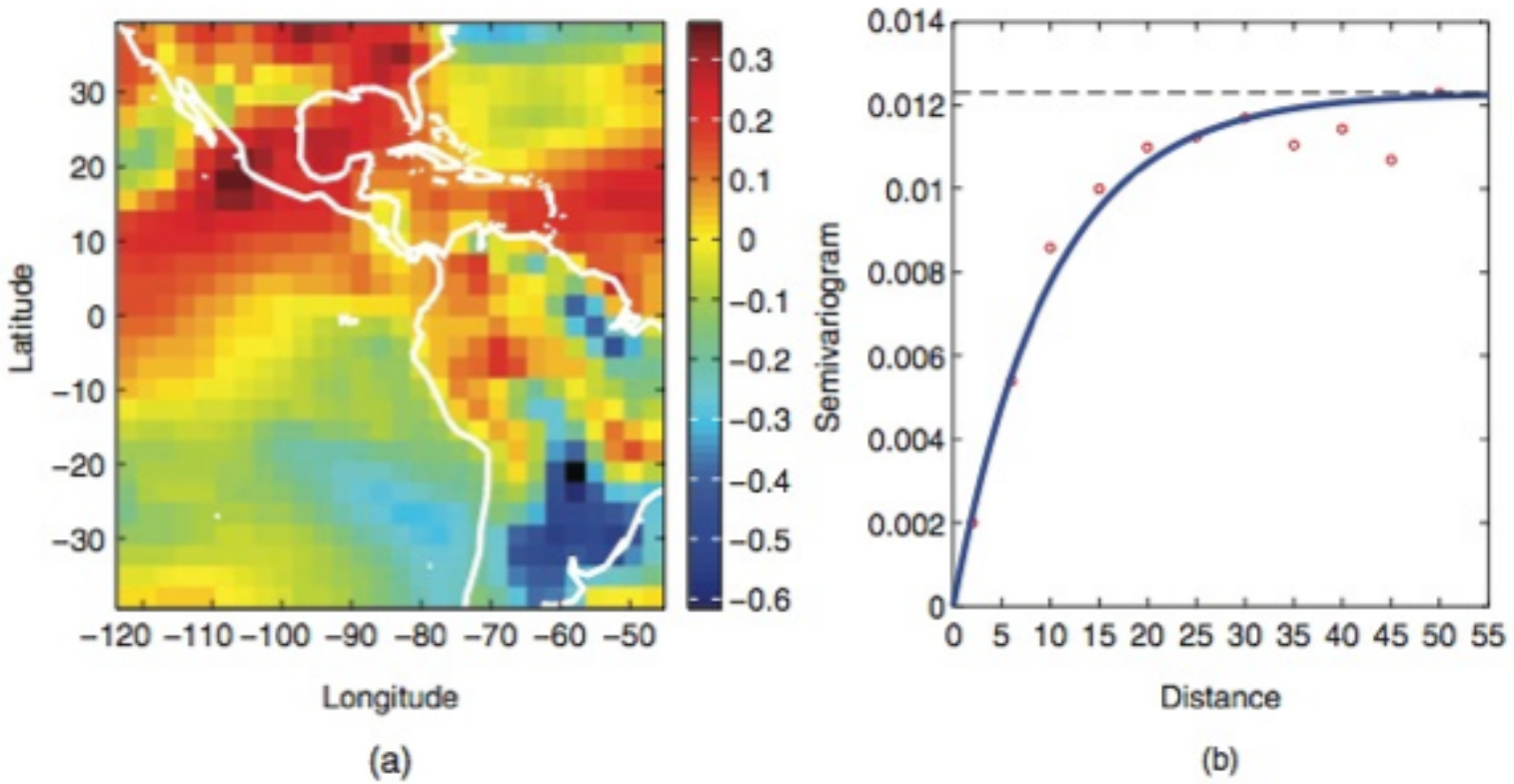

Figure 4.5 **(a)** Temperature change (1990s minus 1980s) in 28 × 28 grid cells over the Americas. The values were originally produced by an NCAR model, and they represent "the truth." **(b)** Empirical and fitted (exponential model) semivariogram after the values in **(a)** were detrended by latitude and longitude.

What about this model?

Using appropriate conditional distributions, the following HM results in non-Gaussian (but continuous) behavior for $Z(\cdot)$.

Data model: Conditional on σ_ε^2, and for $i = 1, \ldots, m$,

$$Z(\mathbf{s}_i)|Y(\mathbf{s}_i), \sigma_\varepsilon^2 \sim ind.\, Gau(Y(\mathbf{s}_i), \sigma_\varepsilon^2)\,.$$

Process model 1: Conditional on $\boldsymbol{\beta}$, $\sigma^2(\cdot)$, and $\rho_Y(\cdot,\cdot)$, $Y(\cdot)$ is a Gaussian process with the following properties:

$E(Y(\cdot)) = \mathbf{x}(\cdot)'\boldsymbol{\beta}$, and we write $\mathrm{cov}(Y(\mathbf{u}), Y(\mathbf{v})) \equiv \sigma(\mathbf{u})\sigma(\mathbf{v})\rho_Y(\mathbf{u}, \mathbf{v})$.

Process model 2: Conditional on $C_\omega(\cdot,\cdot)$, $\sigma(\cdot)$ is a log Gaussian process with the following properties:

$E(\sigma(\mathbf{s})) \equiv 1$, and we write $\mathrm{cov}(\sigma(\mathbf{u}), \sigma(\mathbf{v})) \equiv \exp(C_\omega(\mathbf{u}, \mathbf{v}) - 1)$, where $C_\omega(\cdot,\cdot)$ is the covariance function for $\omega(\cdot) \equiv \log\sigma(\cdot)$.

Notice that $\omega(\cdot) \equiv \log\sigma(\cdot)$ is a Gaussian process, and if we put $E(\omega(\mathbf{s})) = (-1/2)C_\omega(\mathbf{s}, \mathbf{s})$, $\mathbf{s} \in D$, then on the exponential scale the process $\sigma(\cdot) \equiv \exp(\omega(\cdot))$ does have mean 1, as specified in process model 2. Hence, if the Gaussian process $\omega(\cdot)$ is chosen to have constant variance σ_ω^2, then $E(\omega(\mathbf{s})) \equiv (-1/2)\sigma_\omega^2$, a constant.