Inference for stochastic processes in environmental science

VI: Some problems in visualization

Peter Guttorp
NRCSE

Outline

Contour lines and their uncertainty
  Gaussian upcrossings
  Ozone data

Displaying and analyzing proportions
  Compositional algebra
  Benthic species composition
Contour plots

How display uncertainty in contour lines?
- Standard error of spatial predictions
- Distance between contour lines
- “Fuzzy” contour lines

Ozone data revisited

Return to the NW US ozone data: 1974 June-August median daily maximum ground level ozone data from New Jersey, New York, Connecticut and Massachusetts
- Contour plot using bilinear interpolation
- Kriging with exponential covariance function and nugget
Kriging
Standard error

Both plots

surface

standard error
A better combination

A Gaussian prediction $\hat{Z}(s)$ falls between contour lines between $a$ and $b$ with probability

$$p(s) = \Pr(a < \hat{Z}(s) < b)$$

$$= \Phi((1 - r)q) - \Phi(-rq)$$

where $q=(b-a)/\sigma$ and $r= (\hat{Z}(s) - a)/(b - a)$

If $q=2$ the probability is at most $2/3$ that a statement about the level of $Z(s)$ is correct (Polfeldt, 1999).

How many contours?

A Gaussian prediction $\hat{Z}(s)$ falls between contour lines between $a$ and $b$ with probability

$$p(s) = \Pr(a < \hat{Z}(s) < b)$$

$$= \Phi((1 - r)q) - \Phi(-rq)$$

where $q=(b-a)/\sigma$ and $r= (\hat{Z}(s) - a)/(b - a)$

If $q=2$ the probability is at most $2/3$ that a statement about the level of $Z(s)$ is correct (Polfeldt, 1999).
Consequences

Points close to contour lines are always very uncertain as to whether they should be above or below the line. If the contour lines are well separated there are high probabilities of correctness in the middle between them.

Example

For the ozone data set default contour lines from S-Plus are at about $q=2$ ($\sigma = 4.2$, $b - a = 10$).

Retaining contours at -25, 0 and 25 yields $q = 6$ for which $p(s)$ is

<table>
<thead>
<tr>
<th>$r$</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(s)$</td>
<td>0.5</td>
<td>.79</td>
<td>.95</td>
<td>.99</td>
<td>.999</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Ozone again

A linear transformation of the coordinate system for the ozone data is intended to improve the stationarity

\[ x' = Ax + b \text{ where } A = \begin{pmatrix} 1 & 0.5 \\ -1.33 & 6.67 \end{pmatrix}, b = \begin{pmatrix} 54.6 \\ -206 \end{pmatrix} \]

For comparison with a later example we remove the three stations with highest vertical coordinates in the transformed coordinate system.
Lindgren & Rychlik (1995)
Isotropic Gaussian random field $\xi(t)$
Observe $x_k = \xi(t_k) + \varepsilon(t_k)$, $k=1,\ldots,n$
Write $\xi = (\xi(t_1), \ldots, \xi(t_n))^T$. Recall that
$m_n(t) = E(\xi(t) | x) = r_n(t)^T \Sigma^{-1}x$ where
$r_n(t) = \text{Cov}(\xi + \varepsilon, \xi(t))$ and $\Sigma = [r(t_i - t_j)] + \sigma^2 I$
Furthermore, $(\xi(t) | x)$ is a Gaussian process with mean $m_n(t)$ and
covariance $r_n(x, t) = \text{Cov}(\xi(s), \xi(t) | x)$
\[= r(s - t) - r_n(s)^T \Sigma^{-1} r_n(t)\]
Contour lines as level crossings

Let $\eta_n(t) = \xi(t) - m_n(t)$. Our best estimate of the level curve at $u$ over a set $A$, given data $x$, solves $m_n(t) = u$, $t \in A$.

We want to make statements about level crossings of $\{\eta_n(t) + m_n(t), t \in A\}$.

Take sections through the surface. A level curve for $m_n(t)$ is the union of the solutions to $m_n(t) = u$ over line segments in $A$. Thus we deal with level crossings for nonstationary Gaussian processes on the line–known entities.

Turning level crossings into confidence sets

Let $\tilde{\mu}_0^+(\tau)$ be the first upcrossing of $u$ on a line from $t_0$ to $\tau$. Then a line segment $\Lambda(I, \alpha) = \{L(t_0, \tau), \tau \in I\}$ is a level $\alpha$ confidence interval if

$$1 - \alpha = \int_{\tau} \tilde{\mu}_0^+(\tau) \, d\tau.$$
Final ozone revisit

50% CB

Compositional data

Vector of proportions
\[ z = (z_1, \ldots, z_k)^T \quad z_i > 0 \quad \sum_{i=1}^{k} z_i = 1 \quad z \in \nabla^{k-1} \]

Proportion of taxes in different categories
Composition of rock samples
Composition of biological populations
Composition of air pollution
The triangle plot

The spider plot

(0.55,0.15,0.30)

(0.40,0.20,0.10,0.05,0.25)
An algebra for compositions

Perturbation: For $\xi, \alpha \in \mathcal{V}^{k-1}$ define

$$\xi \oplus \alpha = \left( \frac{\xi_1 \alpha_1}{\sum_1^k \xi_i \alpha_i}, \ldots, \frac{\xi_k \alpha_k}{\sum_1^k \xi_i \alpha_i} \right) \in \mathcal{V}^{k-1}$$

The composition $1 = \left( \frac{1}{k}, \ldots, \frac{1}{k} \right)$ acts as a zero, so $\xi \oplus 1 = \xi$.

Set $\xi^{-1} = \left( \frac{1}{\xi_1}, \ldots, \frac{1}{\xi_k} \right)$ so $\xi \oplus \xi^{-1} = 1$.

Finally define $\xi - \eta = \xi \oplus \eta^{-1}$.

The logistic normal

If $\text{alr}(z) = \left( \log \frac{z_1}{z_k}, \ldots, \log \frac{z_{k-1}}{z_k} \right)^T \sim \text{MVN}(\mu, \Sigma)$

we say that $z$ is logistic normal, in short $Z \sim \text{LN}(\mu, \Sigma)$.

Other distributions on the simplex:

- Dirichlet — ratios of independent gammas
- “Danish” — ratios of independent inverse Gaussian

Both have very limited correlation structure.
Scalar multiplication

Let $a$ be a scalar. Define

$$
\xi \otimes a = \left( \frac{\xi_1^a}{\sum \xi_i^a}, \ldots, \frac{\xi_k^a}{\sum \xi_i^a} \right)
$$

$(\mathbb{R}^k, \oplus, \otimes)$ is a complete inner product space, with inner product given, e.g., by

$$
\langle \xi, \eta \rangle = \text{alr}(\xi) \text{T} N^{-1} \text{alr}(\eta),
$$

where $N$ is the multinomial covariance structure $N = I + jj^T$, where $j$ is a vector of $k-1$ ones.

$\| \xi \| = \langle \xi, \xi \rangle$ is a norm on the simplex.

The inner product and norm are invariant to permutations of the components of the composition.

Some models

Measurement error:

$$
z_j = \xi \oplus \epsilon_j \text{ where } \epsilon_j \sim \text{LN}(0, \Sigma).
$$

Regression:

$$
\xi_j = \xi \oplus \gamma \otimes u_j \text{ centered covariate compositions}
$$

Correspondence in Euclidean space:

$$
\mu_j = \beta_0 + \beta_1 (x_j - \bar{x})
$$

$$
\text{alr}^{-1}(\mu_j) = \text{alr}^{-1}(\beta_0) \oplus \text{alr}^{-1}(\beta_1) \otimes (x_j - \bar{x})
$$

$$
\xi_j \quad \xi \quad \gamma \quad u_j
$$
A regression example

A source receptor model

Observe relative concentration $Y_i$ of $k$ species at a location over time. Consider $p$ sources with chemical profiles $\theta_i$. Let $a_i$ be the vector of mixing proportions of the different sources at the receptor on day $i$.

$EY_i = \sum_{l=1}^{p} a_{lj} \theta_l = \Theta a_i$

$Y = \Theta a_i \oplus \epsilon_i$

$\Theta \sim LN, a_i \sim \text{indep LN}, \epsilon_i \sim \text{zero mean LN}$
Juneau air quality

50 observations of relative mass of 5 chemical species. Goal: determine the contribution of wood smoke to local pollution load.

Prior specification:
\[
\pi(\Theta, \alpha_1, \epsilon_1, \mu, \Gamma, \Sigma_e) = \
\pi(\alpha_1 | \mu, \Gamma) \pi(\epsilon_1 | \Sigma_e) \pi(\mu | \alpha) \pi(\Gamma) \pi(\Sigma_e) \pi(\Theta)
\]

Inference by MCMC.

Wood smoke contribution

95% CL

50% CL
Source profiles

References


