The (un)reliability of contour curves
Excursion sets and contour uncertainty regions

David Bolin
Chalmers University of Technology

joint work with Finn Lindgren

Búzios, RJ, Brazil,
June 18, 2014
PM$_{10}$ in Piemonte: Where is PM$_{10} > 50$?
PM$_{10}$ in Piemonte: Where is PM$_{10}$ > 50? Uncertainty?
The problem setting

We have observations $y = (y_1, \ldots, y_n)$ at locations $(s_1, \ldots, s_n)$ of a latent random field $x(s)$. The model is specified through

- The (possibly non-gaussian) likelihood $\pi(y_i| x(s_i), \theta)$.
- A random field model for $x(s)$, typically including covariates.
- Prior distributions for the parameters.

We estimate the parameters and the posteriors (e.g. using INLA) and use the posterior mean $E(x(s)|y)$ as a point estimate of the latent field.

We are interested in the uncertainty of contour curves and excursion sets for $x(s)|y$.

Later, we will assume that $x(s)$ is Gaussian, so that we are in the LGM framework where INLA can be used for estimation.
Lindgren, Rychlik (1995): *How reliable are contour curves?*  
*Confidence sets for level contours*, Bernoulli

- Regions with a single expected crossing
- Method assumes Gaussian likelihood.
- The confidence band is not simultaneous.
Polfeldt (1999), *On the quality of contour maps*, Environmetrics

- How many contour curves should one use in a contour map?
- Based on calculating the marginal probabilities for the field staying between upper and lower contour levels.
- Method assumes Gaussian likelihood.
- Method does not take spatial dependency into account.
A contour curve of a reconstructed field can (almost) be found from the pointwise marginal distributions.

The uncertainty depends on the full joint distribution.

A credible contour region is a region where the field transitions from being clearly below, to being clearly above.

An excursion region is a region where the field is clearly above (or below) a given level.

Finding excursion regions is closely related to multiple testing.

Solving the problem for excursions solves it for contours.

We now need to

- Give precise definitions for the uncertainty regions.
- Construct a method for finding the regions.
Definitions for functions

**Excursion sets for functions**

Given a function $f(s)$, $s \in \Omega$, the positive and negative excursion sets for a level $u$ are

$$A^+_u(f) = \{ s \in \Omega ; f(s) > u \} \quad \text{and} \quad A^-_u(f) = \{ s \in \Omega ; f(s) < u \}.$$  

**Contour sets for functions**

Given a function $f(s)$, $s \in \Omega$, the contour set $A^c_u$ for a level $u$ is

$$A^c_u(f) = \left( A^+_u(f)^o \cup A^-_u(f)^o \right)^c$$

where $A^o$ is the interior and $A^c$ the complement of the set $A$. 
Excursion sets for random fields

Let $x(s), s \in \Omega$ be a random process. The positive and negative level $u$ excursion sets with probability $1 - \alpha$ are

$$E_{u, \alpha}^+(x) = \arg \max_D \{|D| : P(D \subseteq A_u^+(x)) \geq 1 - \alpha\}.$$  
$$E_{u, \alpha}^-(x) = \arg \max_D \{|D| : P(D \subseteq A_u^-(x)) \geq 1 - \alpha\}.$$  

- $E_{u, \alpha}^+(x)$ is the largest set so that, with probability $1 - \alpha$, the level $u$ is exceeded at all locations in the set.
- Another possible definition of an excursion set would be a set that contains all excursions with probability $1 - \alpha$. This set is given by $E_{u, \alpha}^-(x)^c$.  

Definitions — Excursion sets
Example 1: Gaussian process with exponential covariance

- Gaussian process with exponential covariance function.
- $\mathcal{E}_{0,0.05}(x)$ is shown in red.
- The grey area contains $\{s : P(x(s) > 0) > 0.95\}$.
- The dark red set is the Bonferroni lower bound.
- The black curve is the kriging estimate of $x(s)$. 

Definitions — Excursion sets

David Bolin
Contour sets

Level avoiding sets

Let $x(s), s \in \Omega$ be a random process. The pair of level $u$ avoiding sets with probability $1 - \alpha$, $(M_{u,\alpha}^+(x), M_{u,\alpha}^-(x))$, is equal to

$$\arg \max \{|D^- \cup D^+| : P(D^- \subseteq A_u^-(x), D^+ \subseteq A_u^+(x)) \geq 1 - \alpha\}.$$ 

Uncertainty region for contour sets

Let $(M_{u,\alpha}^+(x), M_{u,\alpha}^-(x))$ be the pair of level avoiding sets. The uncertainty region for the contour set of level $u$ is then

$$E_{u,\alpha}^c(x) = (M_{u,\alpha}^+(x)^o \cup M_{u,\alpha}^-(x)^o)^c.$$ 

- $E_{u,\alpha}^c$ is the smallest set such that with probability $1 - \alpha$ all level $u$ crossings of $x$ are in the set.
Example 2: Gaussian Matérn field

- Gaussian Matérn field measured under Gaussian noise.
- Left panel shows the kriging estimate, in the right panel $E_{0,0.05}^c(x)$ is superimposed in grey.
- The complement of $E_{u,\alpha}^c$ is the union of the pair of level avoiding sets.
Excursion functions

The set $E_{u,\alpha}^+(x)$ does not provide any information about the locations not contained in the set.

We want a visual tool similar to $p$-values (i.e. marginal probabilities), but which can be interpreted simultaneously.

<table>
<thead>
<tr>
<th>Excursion functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The positive and negative $u$ excursion functions, contour avoidance functions and the contour function are defined as</td>
</tr>
</tbody>
</table>

$$F_u^+(s) = \sup\{1 - \alpha; s \in E_{u,\alpha}^+\}, \quad F_u^-(s) = \sup\{1 - \alpha; s \in E_{u,\alpha}^-\},$$  
$$F_u^c(s) = \sup\{1 - \alpha; s \in E_{u,\alpha}^c\},$$  
$$F_u^*(s) = \sup\{\alpha; s \in E_{u,\alpha}^c\}. $$

Each set $E_{u,\alpha}^*$ can be retrieved as the $1 - \alpha$ excursion set of the function $F_u^*(s)$. 

Definitions — Excursion functions
Example 1 (cont): Excursion functions

• $E_{u,\alpha}^+$ is retrieved as the $1 - \alpha$ excursion set of $F_u^+(s)$.

• If the function takes a value close to one, the process likely exceeds the level at that location.

• If the value of the function is close to zero, it is more unlikely that the process exceeds the level at that location.
Outline

Introduction
  Piemonte
  Contours
Definitions
  Excursion sets
  Contour sets
  Excursion functions
Calculations
  Intro
  Parametric families
  Integration
  Latent Gaussian
Application
  Piemonte
  Further examples
Remarks
• There are, in principle, two main problems that have to be solved in order to find the excursion sets.
  1 Probability calculation: e.g. calculate the probability $P(D \subseteq A^+_u(x))$ for a given set $D$.
  2 Shape optimization: find the largest region $D$ satisfying the required probability constraint.

• In practice it may not be computationally feasible to solve the problems separately since the probability calculation requires integration of the joint posterior density.

• We need a method that minimizes the number of probability calculations.

• One way of doing this is to use a parametric family for the possible excursion sets.
The parametric families are based on the marginal quantiles of $x(s)$, $P(x(s) \leq q_\rho(s)) = \rho$, which are easy to calculate.

### One-parameter family

Let $q_\rho(s)$ be the marginal quantiles for $x(s)$, then a one-parameter family for the positive and negative $u$ excursion sets is given by

$$D_1^+(\rho) = \{s; P(x(s) > u) \geq 1 - \rho\} = A_u^+(q_\rho),$$
$$D_1^-(\rho) = \{s; P(x(s) < u) \geq 1 - \rho\} = A_u^-(q_1 - \rho).$$

- Using this parametric family reduces the complexity of the shape optimization to finding the correct value of $\rho$.
- Important: $D_1^*(\rho_1) \subseteq D_1^*(\rho_2)$ if $\rho_1 < \rho_2$.
- This simple one-parameter family can be extended in a number of ways, e.g. by smoothing the marginal quantiles.
Gaussian integrals

- For a Gaussian vector $\mathbf{x}$, the probabilities $P(D \subseteq A^+_u(x))$, $P(D \subseteq A^-_u(x))$, and $P(D^+ \subseteq A^+_u(x), D^- \subseteq A^-_u(x))$ can all be written on the form

$$I(a, b, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_{a \leq \mathbf{x} \leq b} \exp\left(-\frac{1}{2} \mathbf{x}^\top \Sigma^{-1} \mathbf{x}\right) d\mathbf{x},$$

- $a$ and $b$ are vectors depending on the mean value of $\mathbf{x}$, the domain $D$, and on $u$.
- There have been considerable research efforts devoted to approximating integrals of this form in recent years\(^1\).
- For GMRFs, we want to use the sparsity of $Q$.
- We use a method based on sequential importance sampling.

\(^1\)A good introduction given in Genz and Bretz (2009), Computation of Multivariate Normal and t Probabilities, Lecture Notes in Statistics, Springer
A sequential Monte-Carlo algorithm

- a GMRF can be viewed as a non-homogeneous AR-process defined backwards in the indices of $x$.
- Let $L$ be the Cholesky factor of $Q$, then

$$x_i | x_{i+1}, \ldots, x_n \sim N \left( \mu_i - \frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji} (x_j - \mu_j), L_{ii}^{-2} \right),$$

- Let $I_i$ be the integral of the last $d - i$ components,

$$I_i = \int_{a_d}^{b_d} \pi(x_d) \int_{a_{d-1}}^{b_{d-1}} \pi(x_{d-1} | x_d) \cdots \int_{a_i}^{b_i} \pi(x_i | x_{i+1:d}) \, dx,$$

- $x_i | x_{i+1:d}$ only depends on the elements in $x_{N_i \cap \{i+1:d\}}$.
- Estimate the integrals using sequential importance sampling.
- In each step $x_j$ is sampled from the truncated Gaussian distribution $1(a_j < x_j < b_j) \pi(x_j | x_{j+1:d}).$
- The importance weights can be updated recursively.
Putting the pieces together

Calculating excursion sets using a one-parameter family

Assume that $\pi(x)$ is Gaussian and that $D(\rho)$ is a parametric family, such that $D(\rho_1) \subseteq D(\rho_2)$ if $\rho_1 < \rho_2$. The following strategy is then used to calculate $E_{u,\alpha}^+$.  

- Choose a suitable (sequential) integration method.
- Reorder the nodes to the order they will be added to the excursion set when the parameter $\rho$ is increased.
- Sequentially add nodes to the set $D$ and in each step update the probability $P(D \subseteq A^+_u(x))$. Stop as soon as this probability falls below $1 - \alpha$.
- $E_{u,\alpha}^+$ is given by the last set $D$ for which $P(D \subseteq A^+_u(x)) \geq 1 - \alpha$. 
Extension to a latent Gaussian setting

- The previous method can only be used in a purely Gaussian setting with known parameters.
- For the more general latent Gaussian setting, the posterior distribution can be written as

\[ \pi(x|y) = \int \pi(x|y, \theta) \pi(\theta|y) \, d\theta, \]

where \( y \) is data and \( \theta \) the parameter vector.
- For Gaussian likelihoods, \( \pi(x|y, \theta) \) is Gaussian.
- There are a number of, more or less complex, ways we can extend the method to the latent Gaussian setting.
- The simplest is to use an empirical Bayes estimator where \( \pi(x|y) \) is replaced with \( \pi_G(x|y, \theta_0) \), a Gaussian approximation at the mode. Two more accurate methods are:
  - Quantile corrections
  - Numerical integration
Quantile Corrections

The QC method is based on modifying the integration limits in the Gaussian integrals based on the marginal posteriors.

- For each $i$, replace the lower limits $a_i$ with
  $$\tilde{a}_i = \sigma_i \Phi^{-1} (1 - P(x_i > a_i | y)),$$
  where $\sigma_i$ is the marginal standard deviation for $x_i | y, \theta_0$ and $\Phi$ denotes the standard Gaussian CDF.

- Similarly, the upper limits $b_i$ are replaced with
  $$\tilde{b}_i = \sigma_i \Phi^{-1} (P(x_i < b_i | y)).$$

- One then has that $P_G(x_i > \tilde{a}_i | y, \theta_0) = P(x_i > a_i | y)$ and $P_G(x_i < \tilde{b}_i | y, \theta_0) = P(x_i < b_i | y)$, where $P_G(\cdot | y, \theta_0)$ denotes the probability calculated under a Gaussian approximation of the posterior $\pi(x | y, \theta_0)$.

- The QC method is exact if the components $x_i$ are independent.
Numerical Integration

In the NI method, one numerically approximates the excursion function as $F_u^\bullet(s) = \sum_{k=1}^{K} \lambda_k F_{u,k}^\bullet(s)$.

- Here $F_{u,k}^\bullet(s)$ is the level $u$ excursion function calculated for the conditional posterior $\pi_G(x \mid y, \theta_k)$ for a fixed parameter configuration $\theta_k$.

- The configurations $\theta_k$ in the hyper parameter space can, for example, be chosen as in the INLA method and the weights $\lambda_k$ are chosen proportional to $\pi(\theta_k \mid y)$.

- Finally, the desired excursion set for a fixed $\alpha$ is retrieved as the excursion set $A_{\alpha}^+(F_u^\bullet)$ of the excursion function.

The NI method is more accurate than the QC method, but requires $K$ times as many calculations.
Air pollution (PM$_{10}$) data

- The limit value fixed by the European directive 2008/50/EC for PM$_{10}$ is $50\mu g/m^3$. The daily mean concentration cannot exceed this value more than 35 days in a year.
- A region where this value is periodically exceeded is the Piemonte region in northern Italy.
- Cameletti et al (2012/13)$^2$ investigated an SPDE/GMRF model for PM$_{10}$ concentration in the region.
- The goal is to analyse exceedance probabilities of the limit value.
- Daily PM$_{10}$ data measured at 24 monitoring stations during 182 days in the period October 2005 - March 2006.

---

$^2$Cameletti, Lindgren, Simpson, and Rue (2012), Spatio-temporal modeling of particulate matter concentration through the SPDE approach, AStA
Model

- The following measurement equation is assumed,

\[ y(s_i, t) = x(s_i, t) + \mathcal{E}(s_i, t), \]

where \( \mathcal{E}(s_i, t) \sim N(0, \sigma^2_E) \) is Gaussian measurement noise, both spatially and temporally uncorrelated.

- \( x(s_i, t) \) is the latent field assumed to be on the form

\[ x(s_i, t) = \sum_{k=1}^{p} z_k(s_i, t) \beta_k + \xi(s_i, t), \]

where the \( p = 9 \) covariates \( z_k \) are used.

- \( \xi \) is assumed to follow first order AR-dynamics in time

\[ \xi(s_i, t) = a\xi(s_i, t - 1) + \omega(s_i, t), \]

where \(|a| < 1\) and \( \omega(s_i, t) \) is a zero-mean temporally independent Gaussian process with spatial Matérn covariances.
Results for January 30, 2006

Spatial reconstruction

Marginal probabilities

Application — Piemonte
Results for January 30, 2006

Marginal probabilities

$F_{50}^+(s)$
Results for January 30, 2006

Contour function $F_{50}^C(s)$

Signed avoidance $\pm F_{50}(s)$
Further examples: Estimating vegetation increase

- Marginally significant trends in green.
- Excursion set $E_{0.05}^+$ in red.
- There has been a vegetation increase in several parts of the region since the drought period in the early 1980s.
Further examples: activation regions in fMRI studies

Joint work with Yue, Lindquist, Lindgren, Simpson, and Rue.
Further examples: estimating bycatch hotspots

Joint work with Godin, Krainski, Worm, Flemming, and Campana.
• Excursion sets and contour uncertainty regions are important in many applications.
• For latent Gaussian models, we can find these quantities efficiently.
• R package excursions, on CRAN:
  \[
  \text{excursions(alpha=0.05, u=0, type="">",}
  \text{mu=field.expectation, Q=precision.matrix)}
  \]
  \[
  \text{excursions.inla(result.inla, ind=candidates,}
  \text{u=0, type=")=", method="NI")}
  \]
• Current and future developments.
  • For excursion sets, compare with other thresholding methods and a sample based method by French and Sain (2013).
  • For contour uncertainty sets, compare with the methods by Lindgren and Rychlik (1995).
  • Combine method with the work by Polfeldt (1999) to make quantitative statements about joint contour map reliability.
• Cameletti, M., Lindgren, F., Simpson, D., and Rue, H.: Spatio-temporal modeling of particulate matter concentration through the SPDE approach; AStA, 2012
• Lindgren, G. and Rychlik, I.: How reliable are contour curves? Confidence sets for level contours, Bernoulli, 1995
• Polfeldt, T.: On the quality of contour maps, Environmetrics, 1999