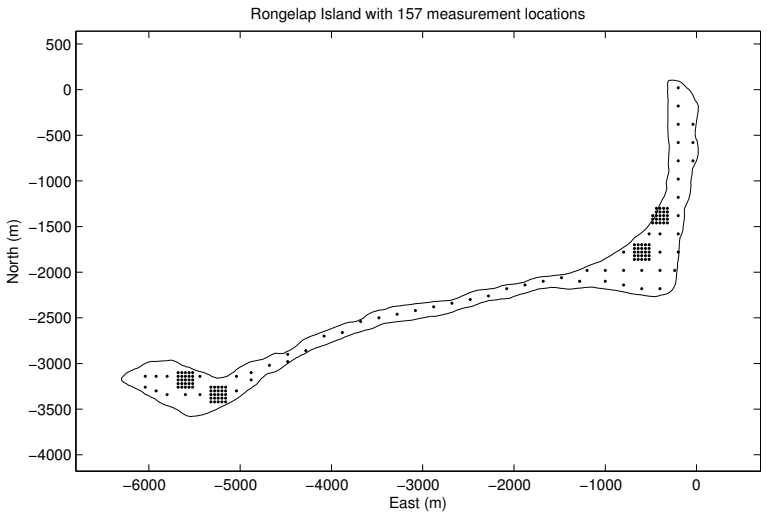
Plan for today

- Recall Latent Gaussian models.
- Prior for parameters of Gaussian process. (We will be Bayesian today.)
- Laplace approximation and numerics for inference, INLA, (Rue et al., 2009)
- INLA shown for geostatistical applications.

Examples of spatial latent Gaussian models

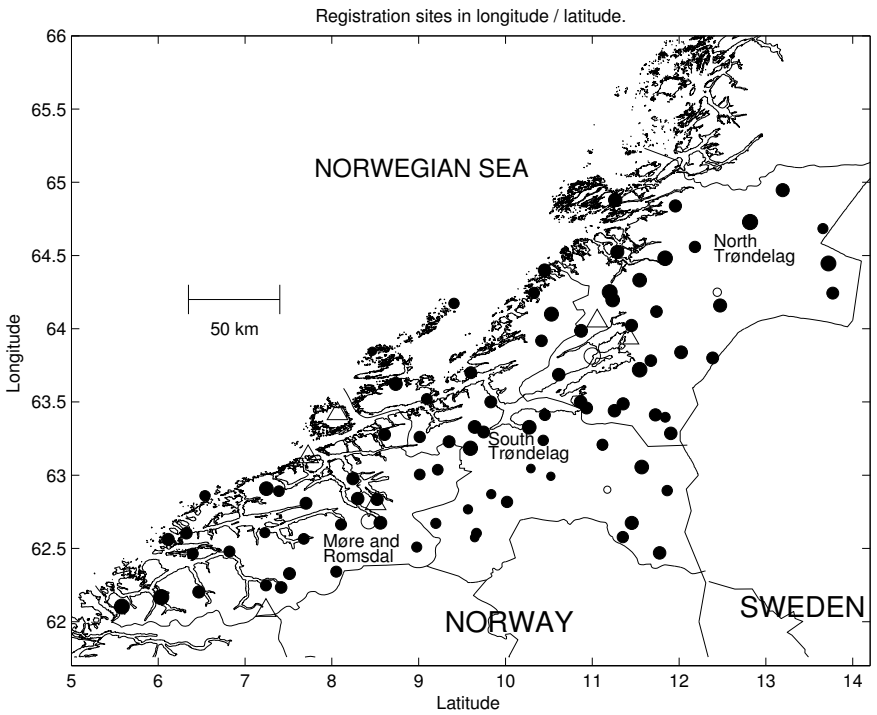
Radioactivity counts: Poisson



Spatial Generalized Linear Model (GLM) : latent log intensity is a GP.

Example of spatial latent Gaussian models

Number of days with rain for $k = 92$ sites in September-October 2006.



Spatial GLM: latent logistic probability is a GP.

Objective

Main goals:

- Fit model parameters (of statistical covariance model in the latent model).
- Predict latent intensity or risk at all spatial sites.

Secondary tasks:

- Outlier detection.
- Spatial design.

Statistical model

Assume the following hierarchical model

1. Observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ where

$$\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\eta}) = \pi(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{j=1}^{n} \pi(y_j \mid x_j)$$

   Often exponential family: Normal, Poisson, binomial, etc.
   $\log \pi(y_j|x_j) = \frac{y_j x_j - b(x_j)}{a(\phi)} + c(\phi, y_j)$. $b(x)$ canonical link.

2. Latent Gaussian process $\boldsymbol{x} = (x_1, \ldots, x_n)$

$$\pi(\boldsymbol{x} \mid \boldsymbol{\eta}) = N[\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\eta})]$$

3. Prior for hyperparameters $\pi(\boldsymbol{\eta})$

NOTE : Last point means we are Bayesian today!

Mixed models - Normal linear case

Common model

- Data model $y_j = \boldsymbol{H}_j\boldsymbol{\beta} + v_j + \epsilon_i$, $\epsilon_j \sim N(0, \tau^2)$
- Prior $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$.
- v_j zero-mean Gaussian random effect having a structured covariance model with parameter $\boldsymbol{\eta}$.
- $x_j = \boldsymbol{H}_j\boldsymbol{\beta} + v_j$

Can integrate out $\boldsymbol{\beta}$.

$$\pi(\boldsymbol{x}|\boldsymbol{\eta}) = \int \pi(\boldsymbol{x}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} = N[\boldsymbol{H}\boldsymbol{\mu}_\beta, \boldsymbol{H}\boldsymbol{\Sigma}_\beta\boldsymbol{H}' + \boldsymbol{\Sigma}(\boldsymbol{\eta})]$$

(We could also augment $\boldsymbol{x}$ with $\boldsymbol{\beta}$ - as long as they are Gaussian it is fine.)
Still a challenge to do inference on $\boldsymbol{\eta}$.

Mixed models - Inference

Common situation that has been hard to infer effectively:

- Frequentist, $\hat{\boldsymbol{\eta}}$: Laplace approximations or estimating equations.
- Bayesian $\pi(\boldsymbol{\eta}|\boldsymbol{y})$: Markov chain Monte Carlo.
- Inference not enough, wish to do model criticism, outlier detection, design, etc. Such goals require fast tools!

Mixed models - GLM

Likelihood is Poisson, binomial, or similar.

- Frequentist: Breslow and Clayton (1993).
- Bayes: Diggle, Tawn and Moyeed (1998), Christensen, Roberts and Sköld (2003), Diggle and Ribeiro (2007).

MCMC - Markov chain Monte Carlo

Around 1990-2000s, MCMC was very popular for Bayesian inference and prediction in latent Gaussian models.
Today MCMC is still very popular, but not so much for latent Gaussian models. Alternatives are Laplace approximations or INLA (Bayes).

Typical MCMC algorithm

Initiate $\boldsymbol{\eta}^1$, $\boldsymbol{x}^1$.
Iterate for $i = 1, \ldots, B$

- Propose $\boldsymbol{\eta}^* | \boldsymbol{x}^i, \boldsymbol{y}$.
- Accept (Set $\boldsymbol{\eta}^{i+1} = \boldsymbol{\eta}^*$) or reject (Set $\boldsymbol{\eta}^{i+1} = \boldsymbol{\eta}^i$), with correct probability (detailed balance).
- For all j; propose $x_j^* | \boldsymbol{x}_{1:j-1}^{i+1}, \boldsymbol{x}_{j+1:k}^{i}, \boldsymbol{\eta}^{i+1}, \boldsymbol{y}$. Accept ($x_j^{i+1} = x_j^*$) or reject ($x_j^{i+1} = x_j^i$).

Pros and cons of such MCMC algorithms

- Converges to sampling from the joint distribution.
- All properties of the distribution can be extracted from MCMC samples.
- Mixing of Markov chain can be very slow. (Blocking or joint proposals help, but also reduces acceptance rate.)
- Gibbs sampler requires conjugate priors. Fast updates, but mixing not better.

Slow mixing means that there is very large autocorrelation in the Markov chain output $(\boldsymbol{x}^1, \boldsymbol{\eta}^1), (\boldsymbol{x}^2, \boldsymbol{\eta}^2) \ldots,$ The method would then need many iterations to cover the distribution properly. Larger moves in the Markov chain reduces the autocorrelation, but tends to have small acceptance rates.

Inference without MCMC sampling

Posterior

$$\pi(\boldsymbol{x}, \boldsymbol{\eta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\eta})\, \pi(\boldsymbol{x} \mid \boldsymbol{\eta})\, \pi(\boldsymbol{y} \mid \boldsymbol{x})$$

In most cases the main tasks are:

- *PREDICTION:* Posterior marginals for x_j, $j = 1, \ldots, n$

$$\pi(x_j \mid \boldsymbol{y})$$

- *PARAMETER ESTIMATION:* Posterior marginals for η_j

$$\pi(\eta_j \mid \boldsymbol{y})$$

Inference

Split the joint density

$$\pi(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{y}) = \pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x}) = \pi(\boldsymbol{y})\pi(\boldsymbol{\eta} \mid \boldsymbol{y})\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})$$

Clearly:

$$\pi(\boldsymbol{\eta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\pi(\boldsymbol{y})\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})} \propto \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})}$$

Marginalization:

$$\pi(\boldsymbol{x}_j \mid \boldsymbol{y}) = \int_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta} \mid \boldsymbol{y})\pi(\boldsymbol{x}_j \mid \boldsymbol{\eta}, \boldsymbol{y}) d\boldsymbol{\eta}$$

Inference

Laplace approximation

$$\hat{\pi}(\boldsymbol{\eta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\hat{\boldsymbol{m}}(\boldsymbol{\eta},\boldsymbol{y})}$$

Use a *Gaussian* approximation $\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})$.
$\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}(\boldsymbol{\eta}, \boldsymbol{y}) = argmax_{\boldsymbol{x}}[\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})]$.

Approximate conjugacy

The Laplace approximation relies on approximate conjugacy. If the full conditional for $\boldsymbol{x}$ is Gaussian, the formula is exact. When we insert a Gaussian approximation at the mode, the approximation depends on the non-Gaussian likelihood. (This cannot be bimodal.)

$$\hat{\pi}(\boldsymbol{\eta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\hat{\boldsymbol{m}}(\boldsymbol{\eta},\boldsymbol{y})}$$

The error of the Laplace approximation (under weak regularity conditions) is *relative* and $\mathcal{O}(n^{-1})$ (Tierney and Kadane, 1986).

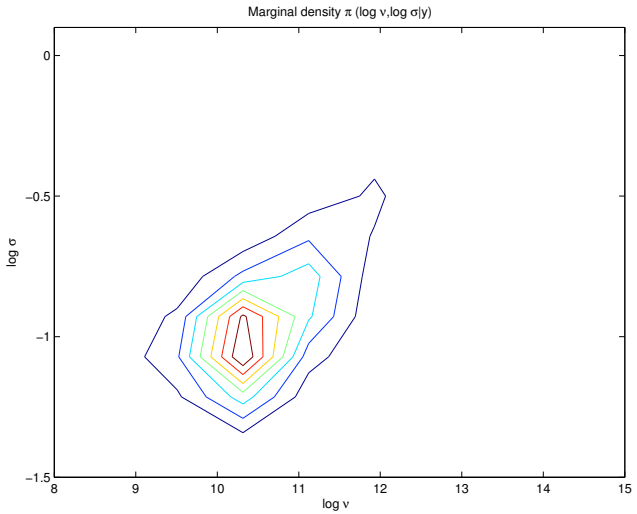
Gaussian approximation of full posterior

$$\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \sum_{j=1}^{n} \log \pi(y_j|x_j)\right)$$

- $\log \pi(y_j|x_j) = \frac{y_j x_j - b(x_{s_j})}{a(\phi)} + c(\phi, y_j)$. $b(x)$ is canonical link.
- Expand GLM part $\log \pi(y_j|x_j)$ to second order.
- Iterative solution to posterior mode $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}(\boldsymbol{\eta}, \boldsymbol{y})$. ('Scoring').
- $\hat{\boldsymbol{m}} = \boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{A}'[\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}' + \boldsymbol{P}]^{-1}(\boldsymbol{z}(\boldsymbol{y}, \hat{\boldsymbol{m}}) - \boldsymbol{A}\boldsymbol{\mu})$.
- Fit Gaussian approximation from Hessian at posterior mode: $\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y}) = N(\hat{\boldsymbol{m}}, \hat{\boldsymbol{V}})$.
- $\boldsymbol{P} = \boldsymbol{P}(\hat{\boldsymbol{m}})$. Size $n \times n$ matrix factorization required.

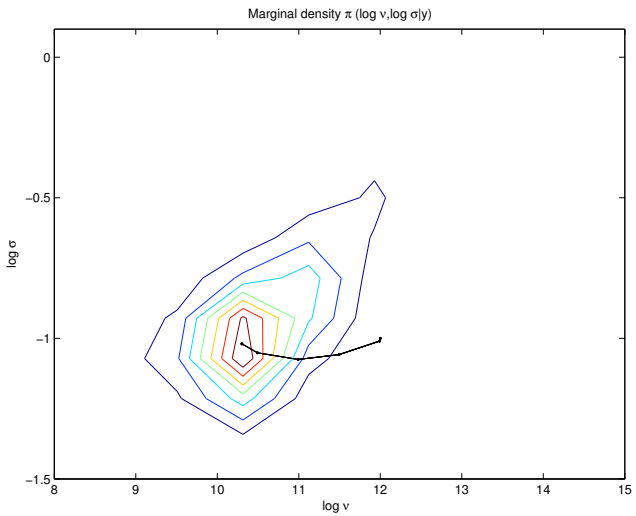
Practical implementation

Numerical approximation of $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$



Practical implementation

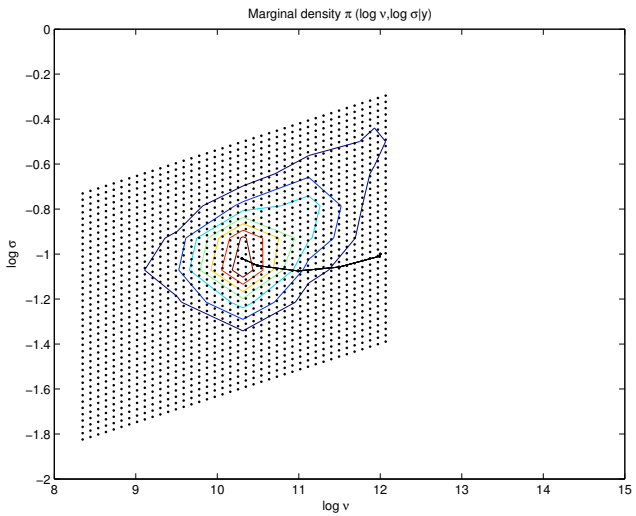
Numerical approximation of $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$, Step 1: Find mode



Each step requires $\boldsymbol{m}(\boldsymbol{\eta}, \boldsymbol{y})$, $\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})$ and Laplace.

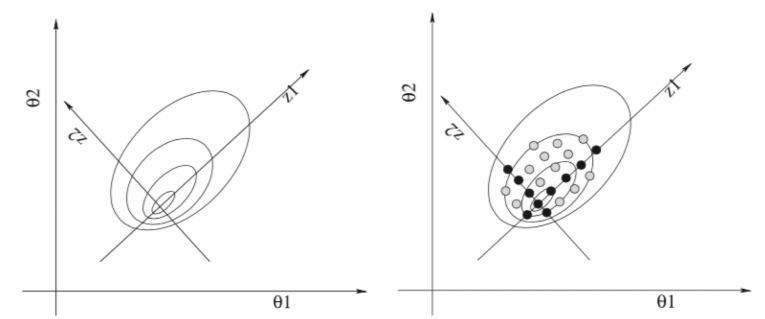
Practical implementation

Numerical approximation of $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$, Step 2: Use Hessian at mode to set grid



Few evaluation points

Sparse numerical approximation of $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$.
Stepping our procedure or central composite design.



Direct approximation of $\pi(x_j|\boldsymbol{y})$

Direct mixture approach for marginal prediction:

$$\hat{\pi}(x_j|\boldsymbol{y}) = \sum_l \hat{\pi}(x_j|\boldsymbol{\eta}_l, \boldsymbol{y})\hat{\pi}(\boldsymbol{\eta}_l|\boldsymbol{y})$$

$\hat{\pi}(x_j|\boldsymbol{\eta}_l, \boldsymbol{y}) = N(\hat{m}_j, \hat{V}_{j,j})$.
$\hat{m}_j = \hat{m}_j(\boldsymbol{\eta}_l, \boldsymbol{y})$, $\hat{V}_{j,j} = \hat{V}_{j,j}(\boldsymbol{\eta}_l, \boldsymbol{y})$.
Element j of posterior mode and j, j of full posterior covariance.

A frequentist solution would just plug in $\hat{\eta}$, the approximate MLE.

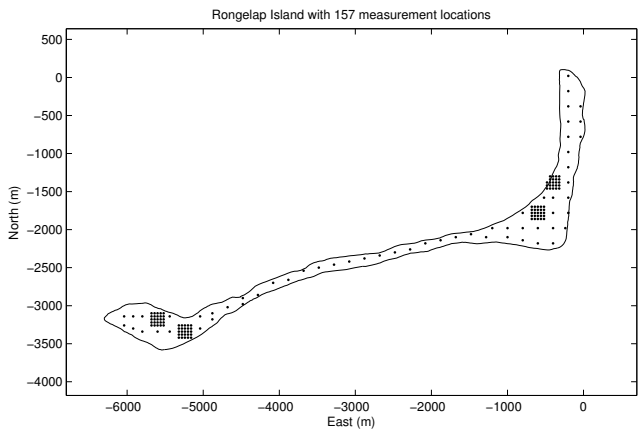
Nested approximation of $\pi(x_j|\boldsymbol{y})$

$$\pi(x_j|\boldsymbol{y},\boldsymbol{\theta}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-j}|x_j,\boldsymbol{y},\boldsymbol{\theta})},$$

Using the Laplace approximation again, for fixed x_j.
$\hat{\pi}(\boldsymbol{x}_{-j}|x_j,\boldsymbol{y},\boldsymbol{\theta})$ approximated by a Gaussian (for each x_j on a grid or design points).

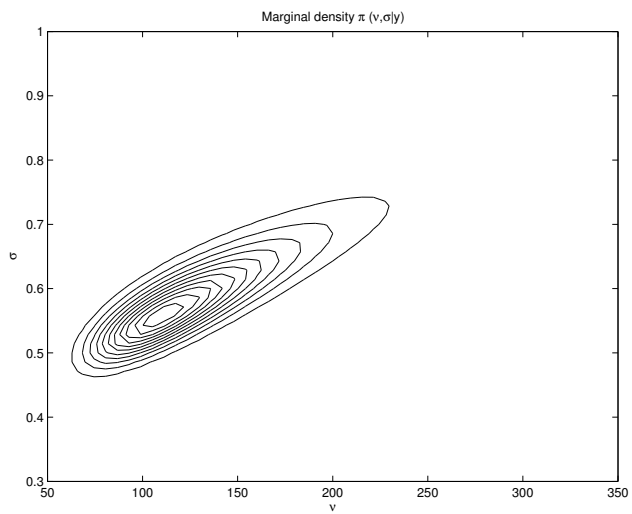
Example: Rongelap

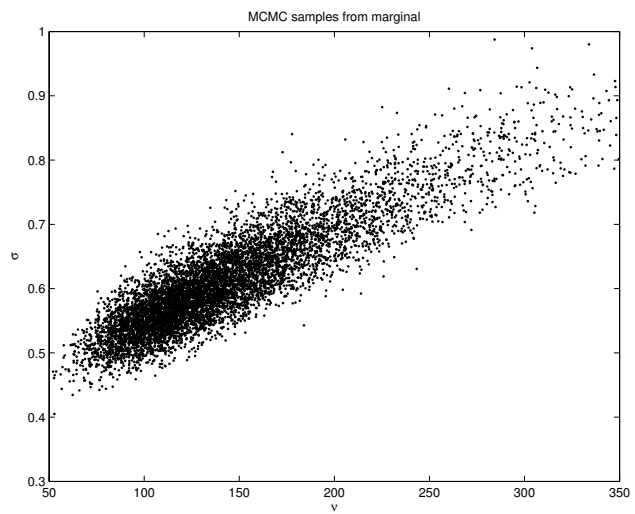
- Radioactivity counts at 157 registration sites. Poisson counts.
- $\boldsymbol{\Sigma}(\boldsymbol{\eta})$ defined from exponential covariance function. $\boldsymbol{\eta} = (\nu, \sigma)$, range and standard deviation.



Marginals for $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$

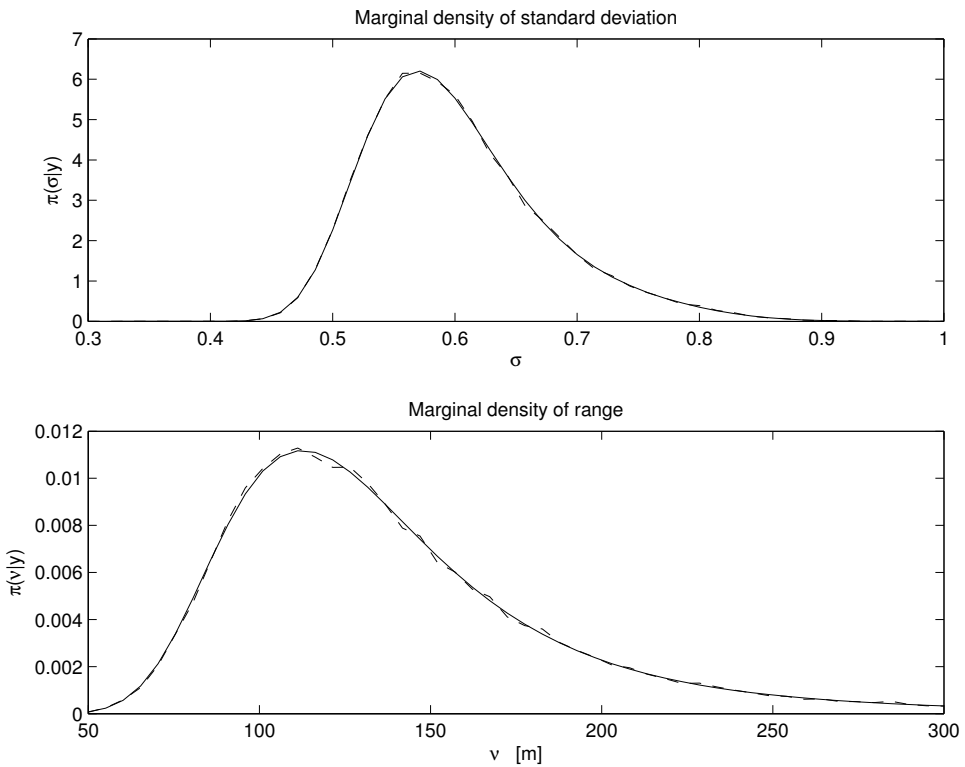
Laplace approximation+numerics (left) and solutions with MCMC (right). Left) Seconds. Right) Minutes.



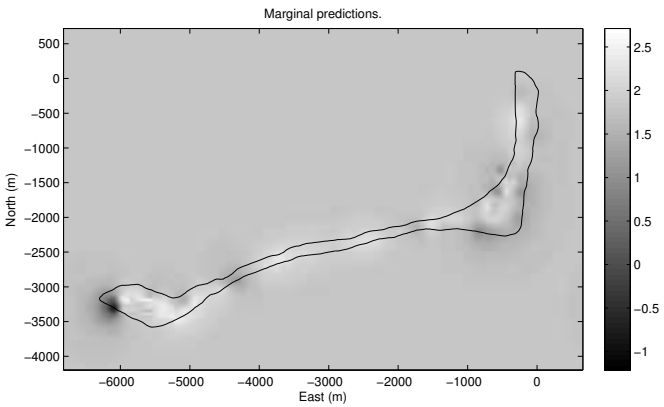


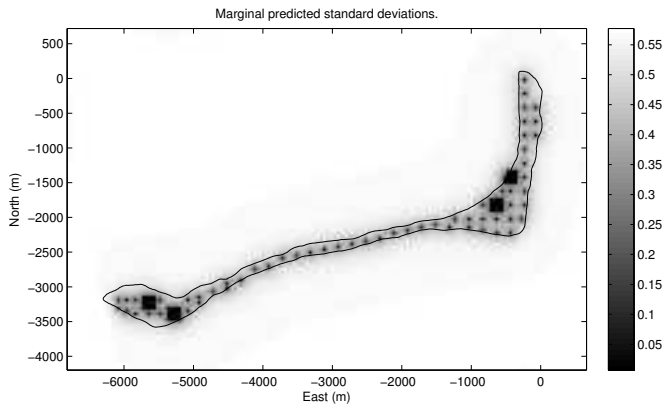
Marginals $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$

Laplace approximation (solid) and MCMC (dashed).



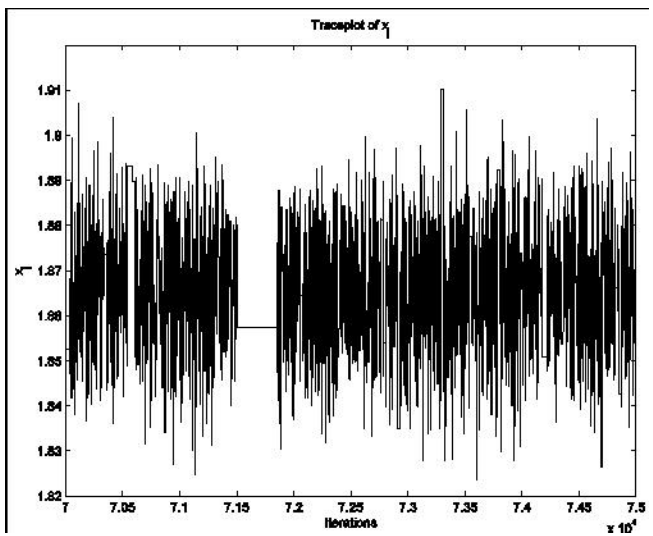
Prediction $\hat{E}(x_j|\boldsymbol{y})$ and $\hat{V}(x_j|\boldsymbol{y})$

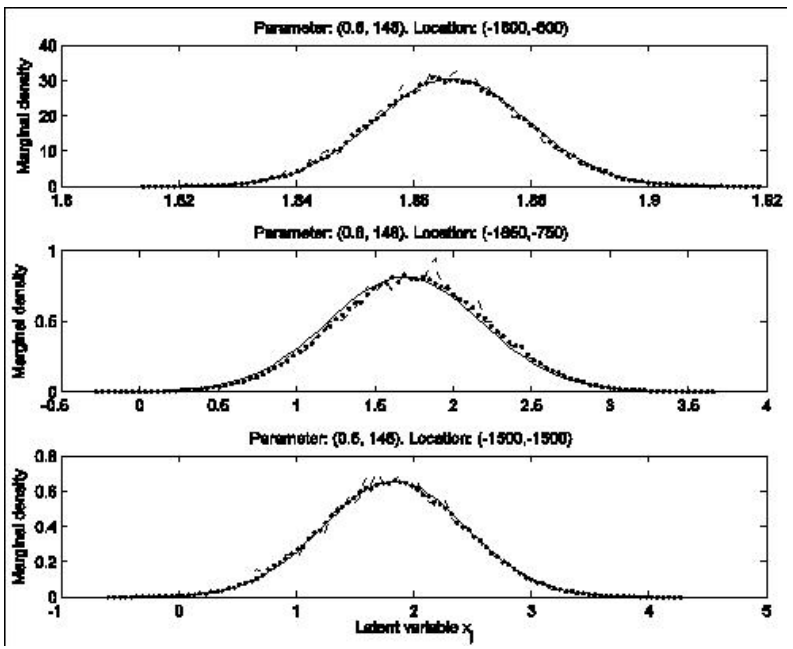




Marginals $\hat{\pi}(\boldsymbol{x}_j|\boldsymbol{\eta}, \boldsymbol{y})$

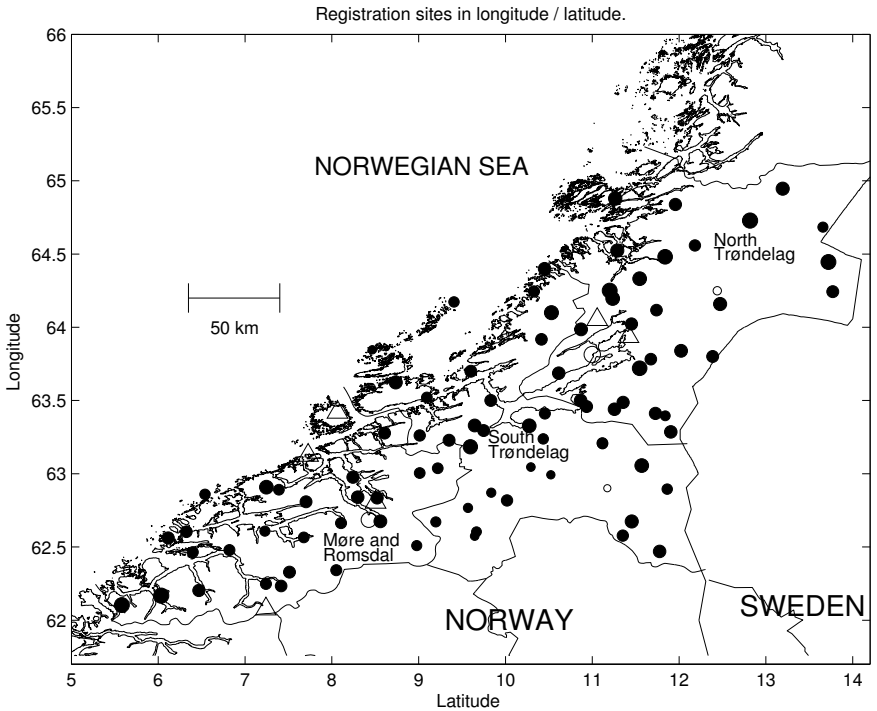
Conditional prediction at one spatial site MCMC (dashed), Importance sampling (dotted) and direct Gaussian approximation (solid).





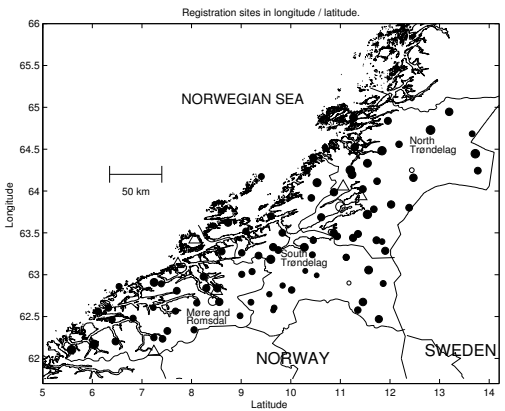
Example: Precipitation in Middle Norway

Number of days with rain for $k = 92$ sites in September-October 2006.



Example: Precipitation in Middle Norway

Binomial data $y_j = \text{Binomial}[\frac{e^{x_j}}{1+e^{x_j}}, 61]$.
Standard GLM gives no significance to East, North, Altitude.
Include only spatial trend.



- Outlier detection
- Spatial design

Outlier detection

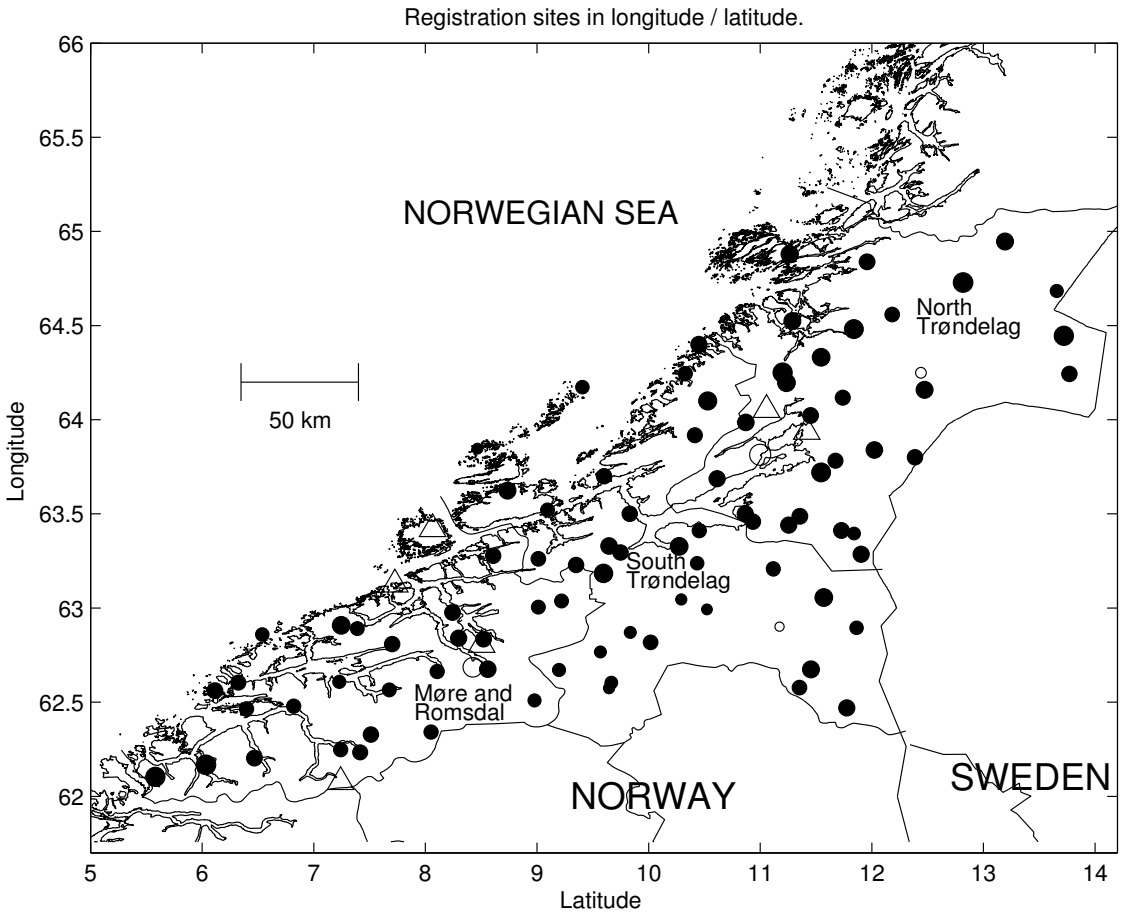
Use crossvalidation $\pi(y_j \mid \boldsymbol{y}_{-j})$.

$$\hat{\pi}(y_j \mid \boldsymbol{y}_{-j}) = \int_{x_j} \sum_l \hat{\pi}(\boldsymbol{\eta}_l \mid \boldsymbol{y}_{-j}) \hat{\pi}(x_j \mid \boldsymbol{\eta}_l, \boldsymbol{y}_{-j}) \pi(y_j \mid x_j) dx_j$$

Inference separately for each y_j. I.e. n times. Approximate predictive percentiles
$\sum_{y_j=0}^{y_{lower}} \hat{\pi}(y_j \mid \boldsymbol{y}_{-j}) = \alpha/2$, $\sum_{y_j=0}^{y_{upper}} \hat{\pi}(y_j \mid \boldsymbol{y}_{-j}) = 1 - \alpha/2$.
Compare (y_{lower}, y_{upper}) with observed y_j.

Results : Outlier detection

Results $\alpha/2 = 0.01$: detect 4 outliers (open circles).



Spatial design

Prospective view: $\boldsymbol{y} \rightarrow (\boldsymbol{y}, \boldsymbol{y}_a)$.
$\boldsymbol{y}_a$ extra data at 'new' spatial registration sites.
'Imagine' these observations - do not acquire them.

Design criterion is: Integrated prediction variance.

$$\hat{I} = \sum_{\boldsymbol{y}_a} \sum_{j} \hat{V}(x_j \mid \boldsymbol{y}, \boldsymbol{y}_a) \hat{\pi}(\boldsymbol{y}_a | \boldsymbol{y})$$

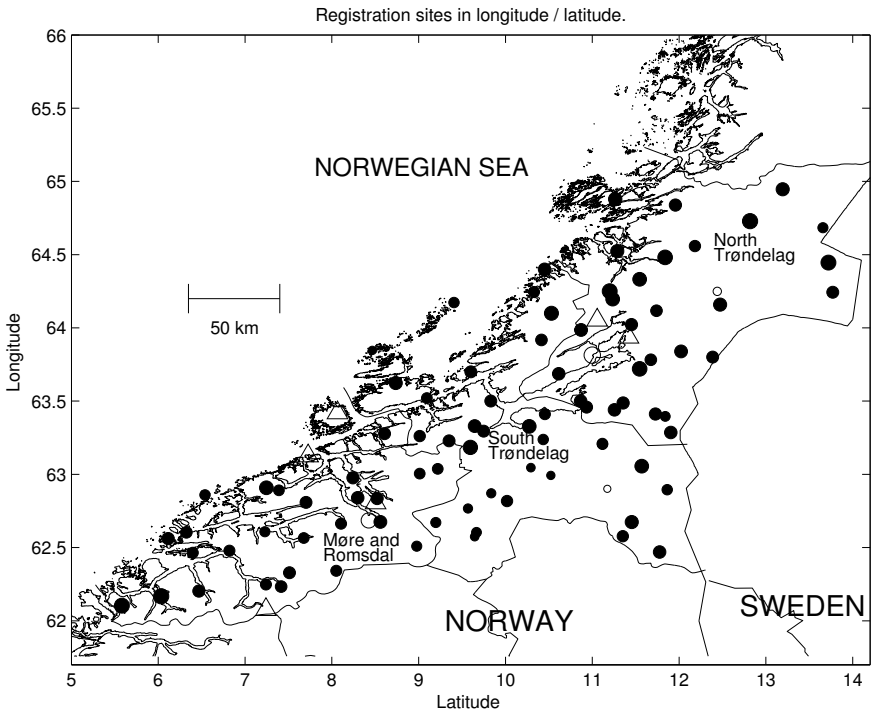
Results: Spatial design

Results of *three* design.

0: Existing design with 88 points (outliers excluded).

A: Currently installed stations, 88 plus 10 known sites (4 outlier sites and 6 sites out of service).

B: 88 plus $10 = 2 \cdot 5$ new random sites around two existing sites (50km radius).



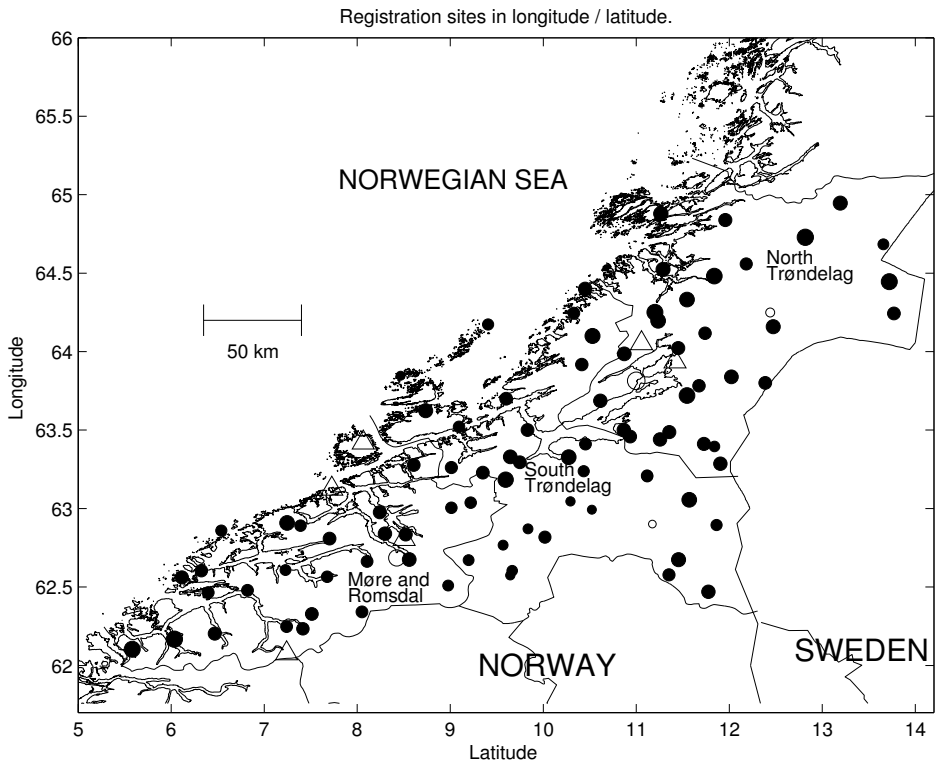
Results: Spatial design

Results of *three* designs.

0: Existing design: $\hat{I}_0 = 18.68$.

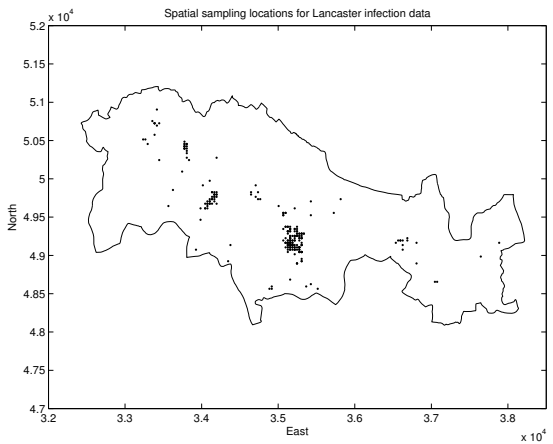
A: Currently installed stations: $\hat{I}_A = 17.94$.

B: Random around two existing sites: $\hat{I}_B = 17.85$



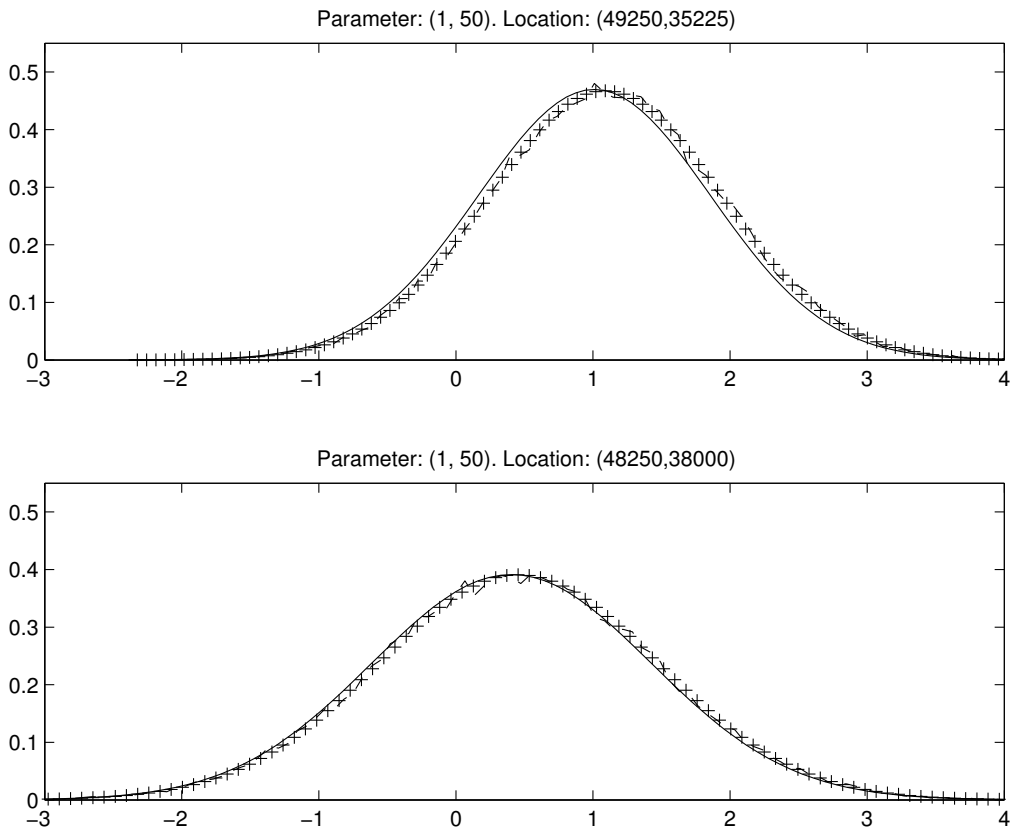
Example: Lancaster disease map

- Number of infections in different regions.
- Binomial data (with small counts).



Example: Lancaster disease map

LA, INLA and MCMC prediction at one site, for two parameter sets.



INLA contribution

- Mixed GLMs with latent Gaussian models cover wide range of applications
- The approximations work well for latent Gaussian models
- Generic routines. Software-friendly. Deterministic results (no Monte Carlo error)
- Enlarge scope of models

Conditions for INLA

- $\dim(\boldsymbol{\eta})$ is not too high
- No. of latent variables $n < 10000$ (Markov assumptions depends on structure). Could turn to approximate GPs.
- Marginals only. Bi-trivariate possible
- Likelihood must be well-behaved, not multimodal.

INLA vs MCMC

MCMC is very general. It explores all aspects of the joint posterior.
Approximate inference (INLA) is much faster. It is tailored to special tasks, such as marginals.
Applicable to much more than spatial data.

INLA software

INLA software: http://www.r-inla.org
Easy to call:
>inla(y x + f(nu,model="iid"), family = c("poisson"), data = data, control.predictor=list(link=1))
Rue et al. (2009)
Routine runs on Gaussian Markov random fields.

(Project Feb 20)