Euroworkshop on Nonparametric models
Schloß Höhenried, November 2001

Bayesian nonparametrics and flexible structured modelling

by Peter Green (University of Bristol, P.J.Green@bristol.ac.uk).

- distributions and dependence
- Dirichlet process and relations
- mixtures
- structured modelling
- space and time

Why nonparametrics?

- “letting the data speak for themselves”

Why Bayesian?

- directness of inference, appealing to non-statisticians
- integrating all sources of uncertainty
- modular: coherent introduction of nonparametric components into structured models
- sequential updating: invariance to permutation
- opportunity of using quantitative prior information if it exists
- uncovering multiple explanations
- most practical and computational objections have been eliminated

Bayesian interpretations of frequentist nonparametric procedures

- smoothing splines
- state-space models
- wavelet thresholding

– not the real focus of contemporary research, but perhaps useful in reminding us of “quasi-Bayesian” character of prior assumptions such as smoothness expressed by a roughness functional.

Bayesian nonparametric modelling of distributions

The basic problem: given observations $Y_1, Y_2, \ldots, Y_n$ from an unknown probability distribution F on a space Ω, make inference about F.

Parametric answer: restrict F to be F_θ for some finite-dimensional parameter θ, place a prior π on θ and use the posterior

$$\pi(\theta|Y) \propto \pi(\theta) \prod_{i=1}^{n} f_\theta(Y_i)$$

Nonparametric answer: only insist that F lies in a bigger (infinite-dimensional?) space, place a prior π on that space, and use the posterior

$$\pi(F|Y) \propto \pi(F) \prod_{i=1}^{n} f(Y_i)$$

Flexible priors on probability distributions

Are there classes of distributions on distributions that are (a) flexible, and (b) permit tractable posterior analysis? A basic ingredient of many of them:

The Dirichlet process

Given a 'base' or 'expectation' probability measure F_0 and a positive scalar parameter c, we write

$$F \sim \mathcal{D}(cF_0)$$

if for every measurable partition $(B_1, B_2, \ldots, B_n)$ of Ω we have

$$(F(B_1), F(B_2), \ldots, F(B_n))$$

$$\sim \text{Dirichlet}(cF_0(B_1), cF_0(B_2), \ldots, cF_0(B_n))$$

Basic properties of the Dirichlet process

$$E(F(B)) = F_0(B)$$

$$\text{var}(F(B)) = \frac{F_0(B)(1 - F_0(B))}{c + 1}$$

so c is a measure of concentration about the base measure F_0.

However, c is also a measure of discreteness. The random F is discrete with probability 1.

If F_0 is continuous, and you draw $F \sim \mathcal{D}(cF_0)$, and then $Y_1, Y_2, \ldots, Y_n | F \sim F$, independently, we find $P(Y_1 = Y_2) = 1/(c + 1)$.

If $c = 0$, then $Y_1 = Y_2 = \cdots = Y_n = Y$ a.s., where $Y \sim F_0$!

If $c = \infty$, then $F = F_0$, and $Y_i \sim F_0$, i.i.d.

Prior to posterior

The beauty of the DP model is the conjugate update:

$$\mathcal{D}(cF_0) + \text{data}(Y_1, Y_2, \ldots, Y_n) = \mathcal{D}(cF_0 + nF_n)$$

where F_n is the empirical distribution of $(Y_1, Y_2, \ldots, Y_n)$.

This is not only of practical benefit, but confers some 'canonical' status on the DP model.

Relatives of the Dirichlet process

The so-called *Mixture of Dirichlet Processes model* (more properly *Dirichlet Process Mixture*) gets round the discreteness problem by introducing 'noise':

$$Y_i|\theta \sim g(\cdot|\theta_i)$$

where

$$\theta_1, \theta_2, \ldots, \theta_n|F \sim F \quad \text{independently}$$

$$\text{and} \quad F \sim \mathcal{D}(cF_0)$$

The conjugacy still helps - Gibbs sampling for the θ_i is trivial - but the inflexibility of the single parameter c for variability remains severe.

Applications of Dirichlet Process Mixtures

By choosing the underlying space Ω, base measure F_0 and data-density g appropriately, an astonishingly wide range of practical statistical methodologies have been devised within this framework - often by West and others, at Duke University.

Often the DPM arises as one ingredient in a fully Bayesian hierarchical model.

- mixture modelling
- nonparametric regression
- autoregression

Connections with finite mixtures

Green and Richardson (*SJS*, 2001) showed and explored a close connection between the MDP model and the finite mixture model

$$Y_i|\theta \sim \sum_{j=1}^{k} w_j\, g(\cdot|\theta_j)$$

where k is random, $\theta_j \sim F_0$, independently,

and $(w_1, w_2, \ldots, w_k) \sim \text{Dirichlet}(\delta, \delta, \ldots, \delta)$.

So far as modelling the Y_i is concerned, the MDP model is just the limit of this as $k \to \infty$ and $k\delta \to c$ (and also according to other limiting regimes). Hardly nonparametric!

Other relatives of the Dirichlet process

- Other neutral-to-the-right processes
- Pólya trees
- Bernoulli trips
- Quantile pyramids
- Dirichlet diffusion trees

See for example Walker, et al., (*JRSS(B)*, 1999), for the 4th, Hjort (*HSSS*, 2002), and for the last, Neal (2001).

Bayesian measurement error modelling

with Sylvia Richardson, Laurent Leblond and Isabelle Jaussent (INSERM, Paris)

Aim: to quantify the association between an outcome Y and a set of covariates X
where covariates are imperfectly observed and only measured through “surrogates”.

Ignoring measurement error and treating the surrogate as the true covariate may produce biased results.

Why be Bayesian here?

- latent covariates with imprecisely specified prior distributions
- combining information on measurement process from several sources
- propagating uncertainty

Model building – structural specifications

- Y known outcome
- X true (latent) covariate
- U observed surrogate for X
- C known covariates

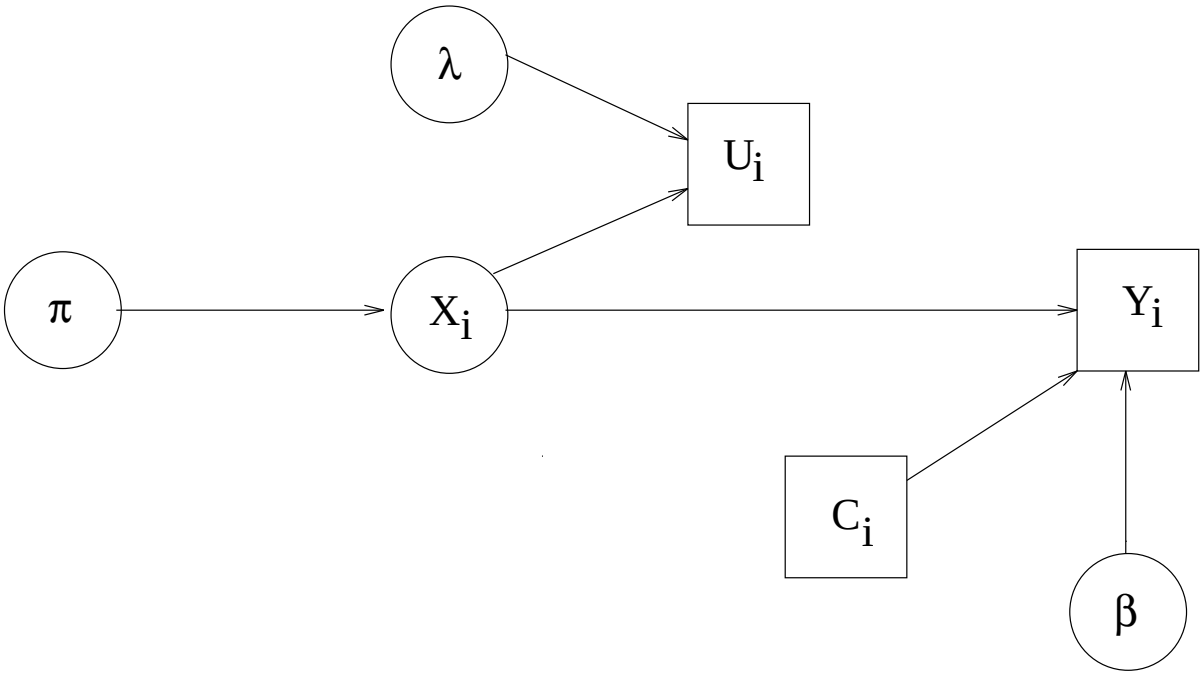
Formulation of local submodels between components using
– conditional independence assumptions
– prior information on the structure of the measurement process

Submodels:

- $p(Y|X, C, \beta)$ regression model
- $p(U|X, \lambda)$ measurement model
- $p(X|\pi)$ prior model

Bayesian analysis using graphical models

Non differential measurement error assumption:
$Y \perp U|X$



Joint distribution:

$$p(\beta)p(\lambda)p(\pi)\prod_i p(X_i|\pi)$$

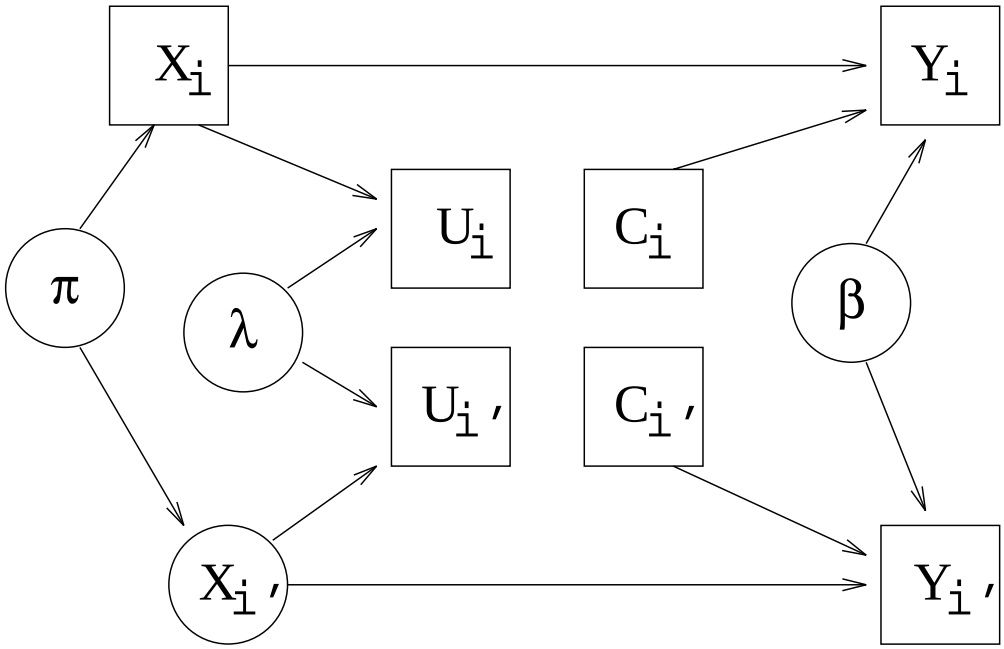
$$\times \prod_i p(U_i|X_i,\lambda)\prod_i p(Y_i|X_i,C_i,\beta)$$

Where does quantitative information on measurement model come from ?

One possibility: design with a validation group:

reference method which can be used to get information on X from a subgroup where both X and U are recorded.

Designs with a validation group



- transfer of information on λ from the validation group to the main study
- strengthens inference about regression parameters β

Problems in specifying prior for $p(X|\pi)$

Some approaches

- pseudo-likelihood *(Carroll, 1993)* based on plugging in an empirical estimate of $p(X|\pi)$ based on the validation subgroup
- non parametric modelling of $p(X|\pi)$ via NPML *(Roeder, Carroll, Lindsay, JASA 1996)*
- joint modelling of $p(X, U|\lambda)$ as a Multivariate normal where λ specified in terms of a Dirichlet Process *(Müller and Roeder, Biometrika, 1997)*
- semi-parametric model for $p(X|\pi)$ via a mixture of gaussian distributions with an unknown number of components

Mixture model for $p(X|\pi)$

$$X_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \quad \text{independently for } i = 1, 2, \ldots, n$$

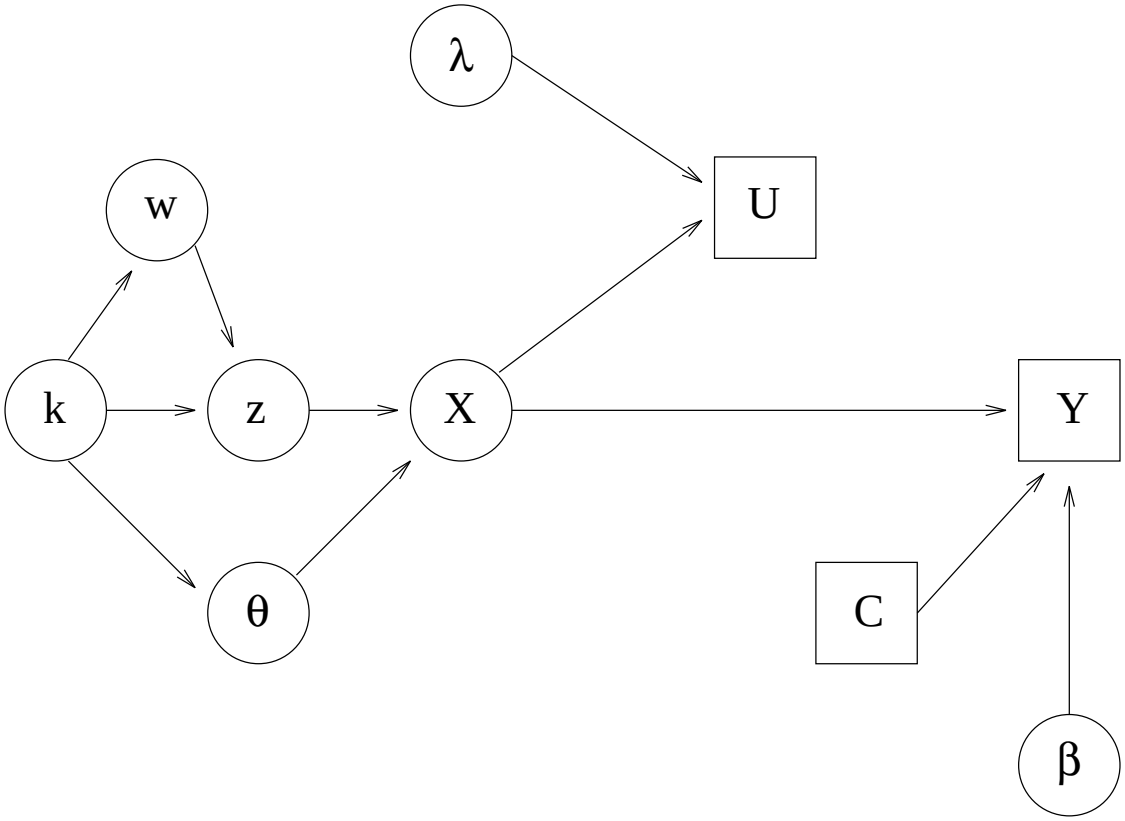
$f(\cdot|\theta)$ is a given parametric family
$\{\theta_j\}, \{w_j\}, k$ **unknown**

The model can be formulated using latent allocation variables:

$$p(z_i = j) = w_j \quad \text{independently for } i = 1, 2, \ldots, n$$

$$X_i|z \sim f(\cdot|\theta_{z_i}) \quad \text{independently for } i = 1, 2, \ldots, n$$

Measurement error model with mixture prior



Of course, computing in such models would be quite impossible by conventional methods.

With MCMC, most of the variables can be updated singly or in small groups, by Gibbs or Metropolis moves.

We update k (with consequent changes to w, z and θ) by reversible jump *split/merge* moves.

Implementation in the case of a logistic regression with validation group design

Prior for X: normal mixture model

$$X \sim \sum_{j=1}^{k} w_j \phi(\cdot|\mu_j, \sigma_j^2), \quad k \text{ unknown}$$

Measurement error: e.g., lognormal

$$\log U_i \sim N(\alpha_0 + \alpha_1 \log X_i, \lambda^{-1})$$

Regression model for disease status: logistic model

$$Y \sim \text{Bernoulli}(\{1 + \exp[-\beta^T(X, C)]\}^{-1})$$

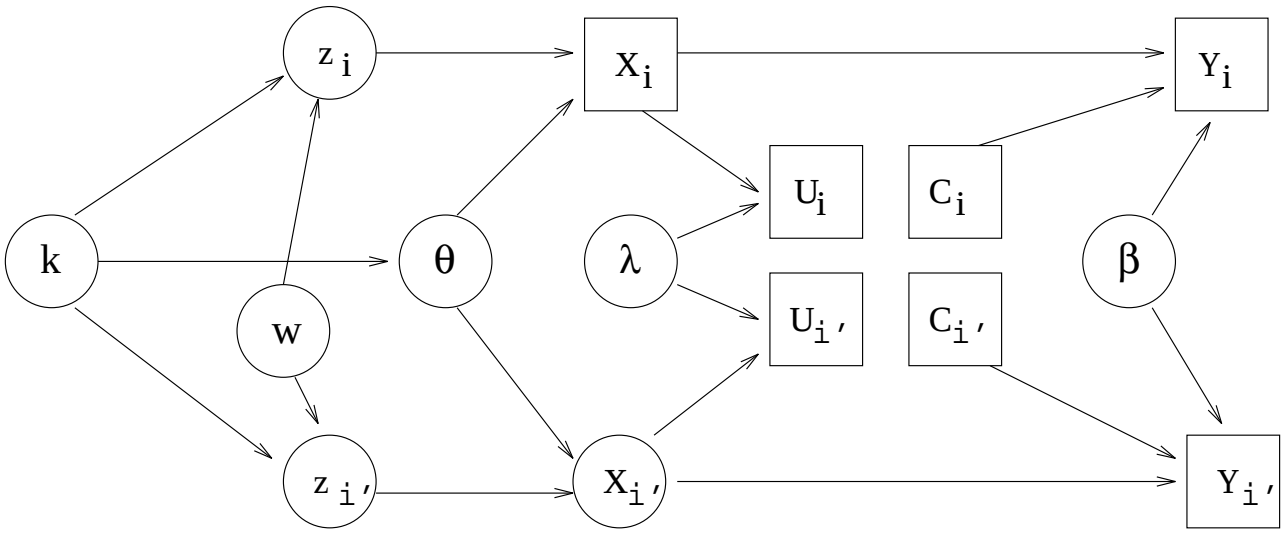


Illustration on a study of the risk of coronary heart disease (CHD) as a function of blood cholesterol

Total cholesterol (TC) and Low density cholesterol (LDL) on 256 subjects: 113 cases, 143 controls.

$\rightarrow$ can we use TC = U as a surrogate for LDL = X?
$\rightarrow$ a validation subgroup with 32 cases and 40 controls is chosen at random

Logistic regressions of CHD on cholesterol level:

- regression on X, complete data set $(n = 256)$

$$\beta_1 = 0.66 \quad (0.34)$$

- regression on U, complete data set $(n = 256)$

$$\beta_1 = 0.54 \quad (0.31)$$

- regression on X, validation group $(n = 72)$

$$\beta_1 = 0.93 \quad (0.68)$$

- Bayesian analysis (validation and main study)

$$\beta_1 = 0.62 \quad (0.44)$$

Performance of mixture priors in measurement error models

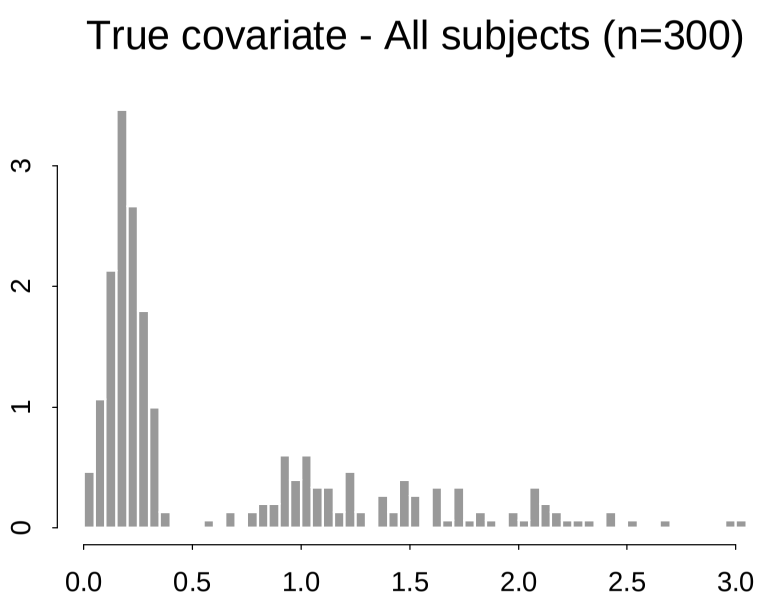
Simulation set up: 50 replications
270 subjects in main study, 30 in validation group.
X drawn from an asymmetric normal mixture :

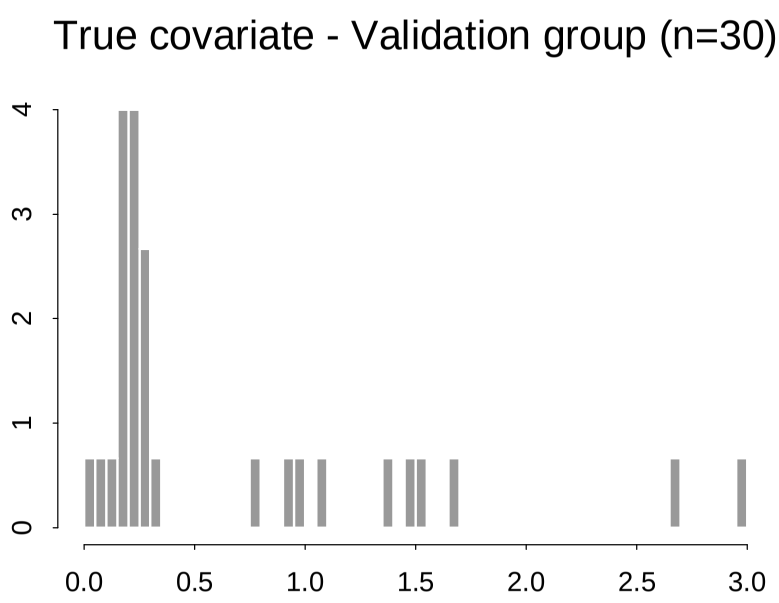
$$0.6N(0.19,(0.08)^2)+0.2N(1.05,(0.2)^2)+0.2N(1.63,(0.48)^2)$$

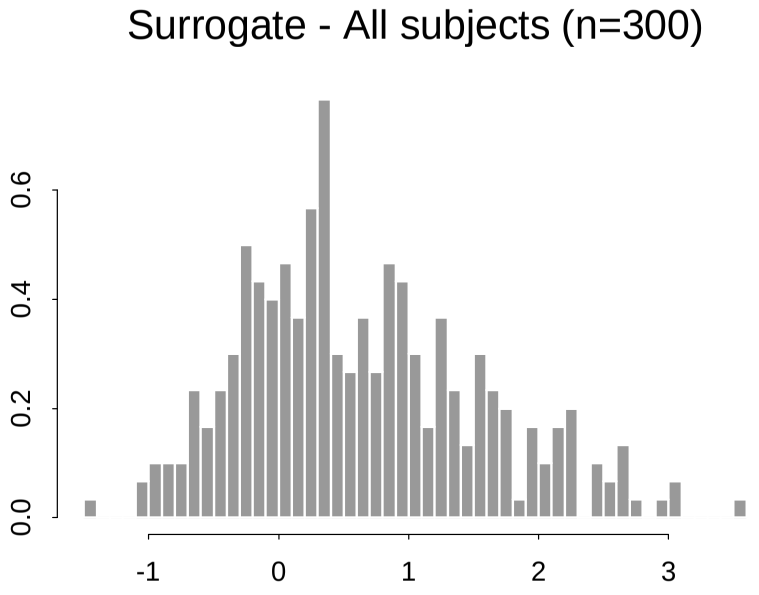
Measurement model : $U \sim N(X, \lambda^{-1})$

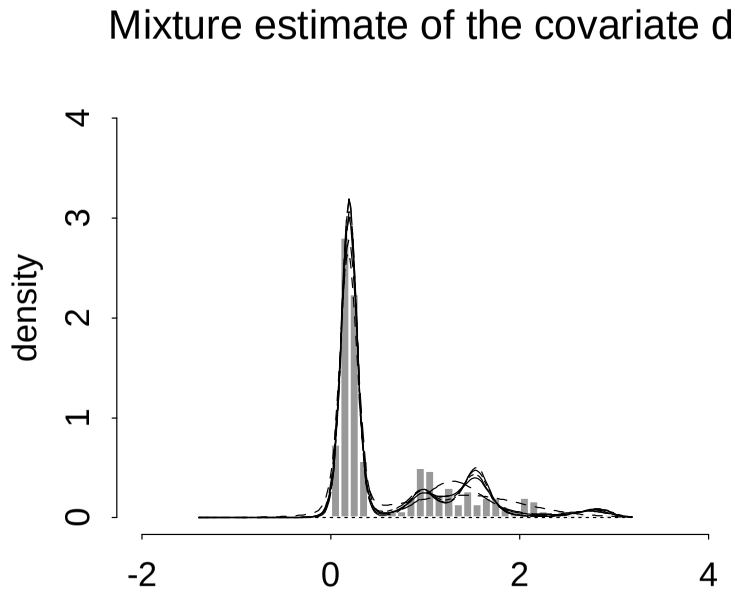
Logistic disease model :
logit $P(Y=1|X) = \beta_0 + \beta_1 X$

		analysis	
	true	mixture prior	gaussian prior
$\overline{\lambda}$	3	2.82 (0.41)	2.37 (0.51)
$\overline{\beta_0}$	-0.8	-0.86 (0.19)	-1.05 (0.27)
$\overline{\beta_1}$	0.4	0.52 (0.25)	0.76 (0.32)
mse(β_1)		0.053	0.092


True covariate - All subjects (n=300)
0
1
2
3
0.0
0.5
1.0
1.5
2.0
2.5
3.0


True covariate - Validation group (n=30)
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Surrogate - All subjects (n=300)
0.0
0.2
0.4
0.6
-1
0
1
2
3


Mixture estimate of the covariate density
density
0
1
2
3
4
-2
0
2
4

Bayesian nonparametric modelling of dependence

Hjort (*HSSS*, 2002) discusses some Bayesian variants on local polynomial regression methods.

Here, however, we focus on highly data-adaptive methods for particular spatial and temporal problems, built on flexible structured models.

Hidden Markov models, spatial mixtures, and disease mapping

(with Sylvia Richardson (INSERM $\rightarrow$ Imperial) and Carmen Fernández (Bristol $\rightarrow$ St. Andrews))

Small area disease mapping

In regions indexed $i = 1, 2, \ldots, n$:
$y_i =$ observed count of disease incidence
$E_i =$ expected count based on population size, adjusted for age and sex, etc.

$y_i / E_i =$ standardised mortality (morbidity) ratio (SMR)

Standard assumption: $y_i \sim \text{Poisson}(\lambda_i E_i)$
$\Rightarrow$ inference on relative risks $\{\lambda_i\}$

Structure of prior for relative risks

Continuously distributed MRF's for the joint distribution of the $\{\lambda_i, i = 1, 2, \ldots, n\}$:
Besag, York and Mollié (1991), Clayton and Bernardinelli (1992), Best, et al (1999), Wakefield and Morris (1999)

Parameters characterising spatial dependence are constant across entire study region

$\Rightarrow$ potential risk of over-smoothing and masking of local discontinuities, due to global effect of the parameters (concern borne out by empirical studies)

Hidden discrete-valued random fields

Common feature of several attempts to address this: replace continuously varying random field for $\{\lambda_i\}$ by an allocation / partition model of the form

$$\lambda_i = \lambda_{z_i}$$

$\{\lambda_j, j = 1, 2, \ldots, k\}$ characterise k components $\{z_i, i = 1, 2, \ldots, n\}$ are *allocation variables* taking values in $\{1, 2, \ldots, k\}$

Moving spatial dependence one level higher in the hierarchy, to the $\{z_i\}$ has the potential for greater spatial adaptivity (again seen empirically).

Discreteness in the prior is not imposed on posterior inference. Under Bayesian model averaging, the posterior mean risk surface can provide a smooth estimate.

Models in this framework

include

- clustering or segmentation models of Knorr-Held and Raßer (2000) and Denison and Holmes (2001)
- Green and Richardson (2000) – Potts model for $\{z_i\}$, with the number of states and strength of interaction unknown (we retain a Markovian structure for the $\{z_i\}$)
- Fernández and Green (2000) – spatial mixture models – spatial dependence is pushed yet one level higher: the $\{z_i\}$ are conditionally independent given weights $w_{ij} = P(z_i = j)$

Hidden Markov model approach

Basic mixture set-up

$$y_i \sim \sum_{j=1}^{k} w_j f(\cdot|\theta_j) \quad \text{independently}$$

$$\equiv$$

introduce latent allocation variables $\{z_i\}$ with

$$\begin{aligned} y_i|z &\sim f(\cdot|\theta_{z_i}) \\ p(z_i = j) &= w_j \end{aligned}$$

Temporal HMM set-up

As above, but i now represents (discrete) time.

Data are a time series (y_i), and (z_i) is now a Markov chain.

Extension to spatial case for disease mapping

Write relative risk as λ_{z_i} in place of λ_i.

$$y_i|z \sim \text{Poisson}(\lambda_{z_i} E_i)$$

where $\{z_i\}$ is a spatially dependent random field with $z_i \in \{1, 2, \ldots, k\}$.

More commonly we would have covariates x_i and use the model:

$$y_i|z \sim \text{Poisson}(\lambda_{z_i} E_i e^{x_i'\beta})$$

Allocation models

In each case, spatial context determined by assumed neighbourhood structure – we say 'adjacent' $\equiv$ 'have common boundary' ($i \sim j$). For rare diseases, more complex dependence not justified.

The formulations we have implemented and explored:

- Potts model: $p(z) = \exp(\psi U(z) - \theta_k(\psi))$ where $U(z) = \#\{i \sim j : z_i = z_j\}$ = number of like-coloured neighbour pairs.
- multinomial allocation – $p(z_i = j) = w_{ij}$ – using either
  - logistic-normal weights: $w_{ij} = \exp(x_{ij}) / \sum_{j'} \exp(x_{ij'})$
  - grouped continuous weights: $w_{ij} = \Psi(x_i - \delta_j) - \Psi(x_i - \delta_{j-1})$

  where (x_{ij}) and (x_i) are Gaussian random fields.

Interpretation and inference in HMRFs and partition models

Do we really believe there are k groups of regions with identical relative risks?

- model is being used in a 'semi-parametric' fashion, not to identify clusters
- inference on $\{\lambda_{z_i}\}$ rather robust to details of prior structure – 'borrows strength' between regions in an adaptive way (by Bayesian model averaging)
- avoid over-smoothing of relative risks
- interpret inference on k and z with caution (diagnostic / exploratory)

Some issues in model choice for spatial epidemiology

- objectives of the model and of the choice
- statistical paradigm
- specific criteria

One key consideration is the extent to which it is believed that all relevant covariates have been measured and included appropriately in the model.

(We can accept that ’all models are wrong’ without accepting that all models are *equally* useless!)

Confounding between spatial structure of covariates and random effects

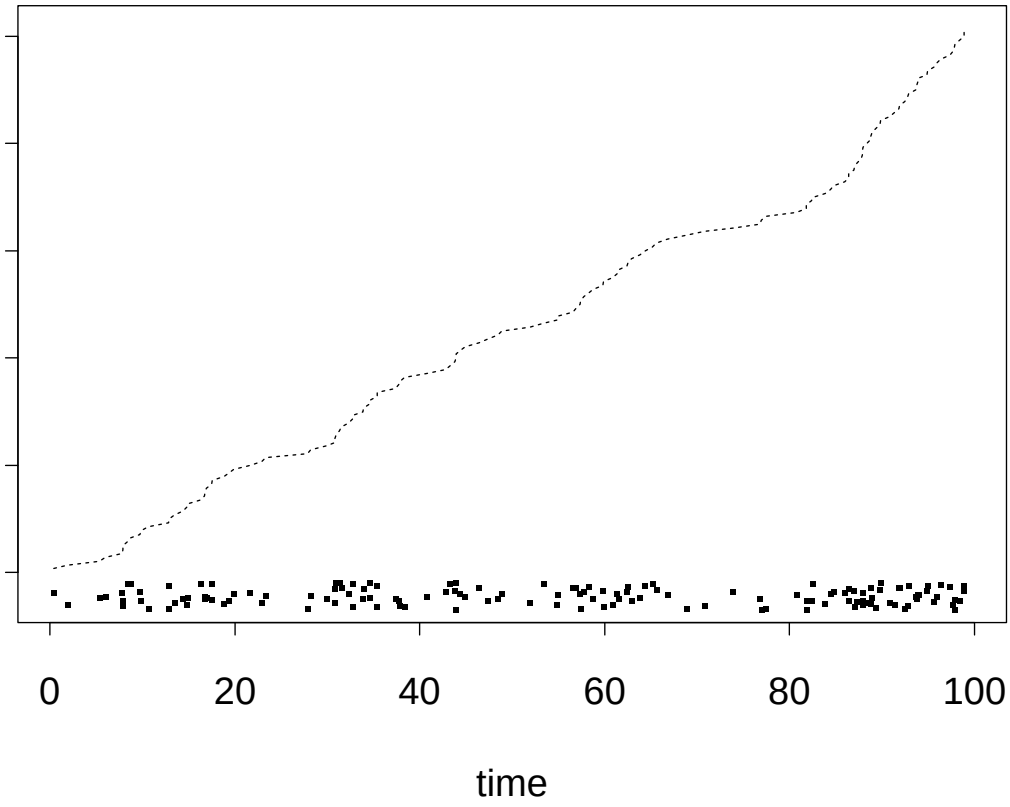
A periodically-voiced concern is over whether fitting flexible spatial models in addition to covariates systematically 'dilutes' estimates of covariate effects (the implication being to be deliberately modest in allowing for unmeasured covariates in order not to eliminate the significance of the measured ones).

This concern is probably unfounded. See the partial report of an on-going simulation study by Richardson (*HSSS*, 2002). If spatial correlation between covariates and random effects is generated, there will be confounding – positive *or* negative bias, otherwise, not.

Multiple change points in point processes

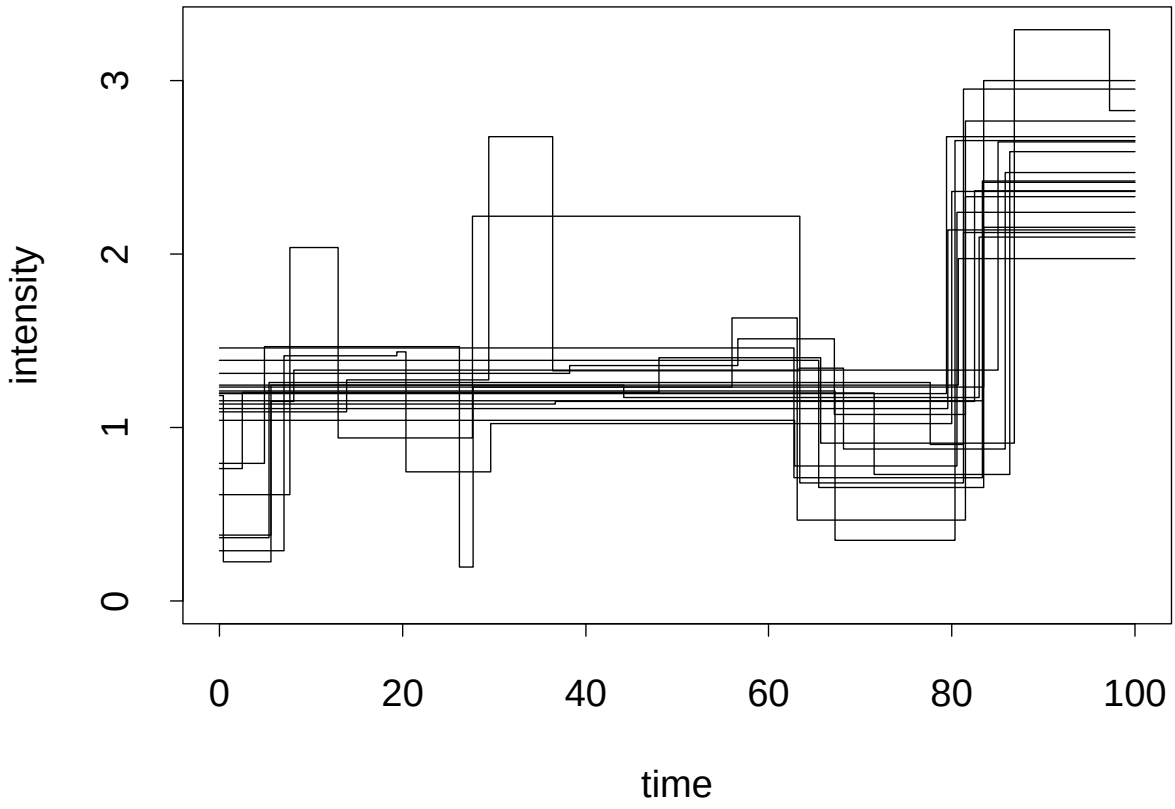
Example:
cyclones hitting the Bay of Bengal

141 cyclones over a period of 100 years
(a cyclone is a storm with winds > 88 km h^{-1}).

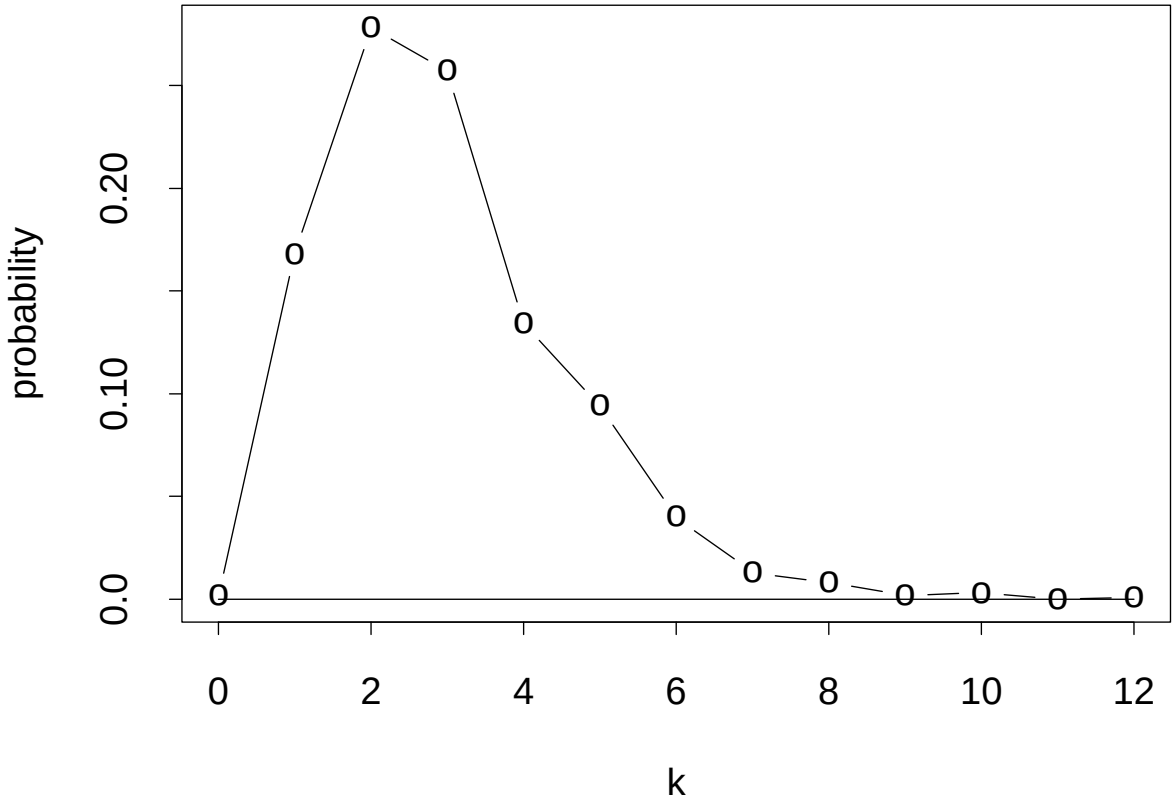


Our model is that the intensity as a function of time is a step function, with an unknown number of steps.

The number of steps k is Poisson(λ), with $\lambda = 3$, the step function positions are drawn from the joint density $\propto s_1(s_2 - s_1)(s_3 - s_2)\ldots(s_k - s_{k-1})(L - s_k)$ and the step heights are independent Gamma(α,β), with $\alpha \sim \Gamma(2, 2)$ and $\beta \sim \Gamma(1, n/L)$.

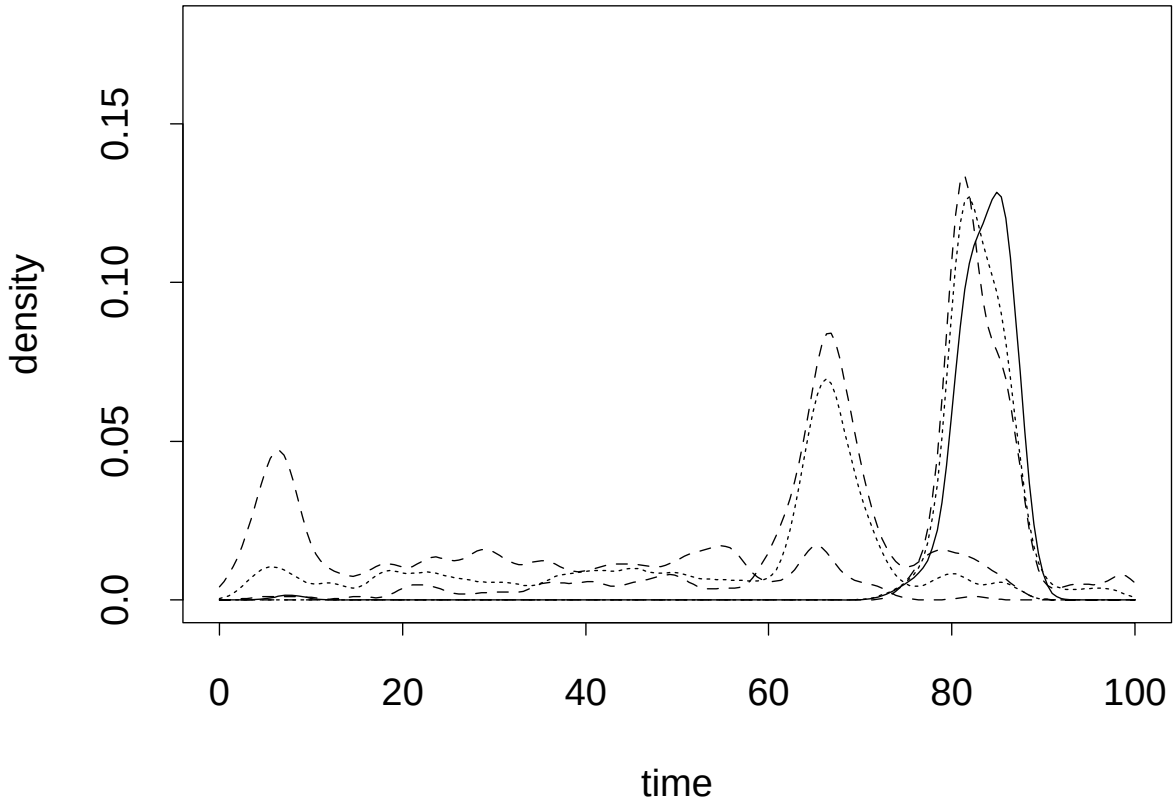

intensity
0
1
2
3
0
20
40
60
80
100
time

Posterior for the number of change points k

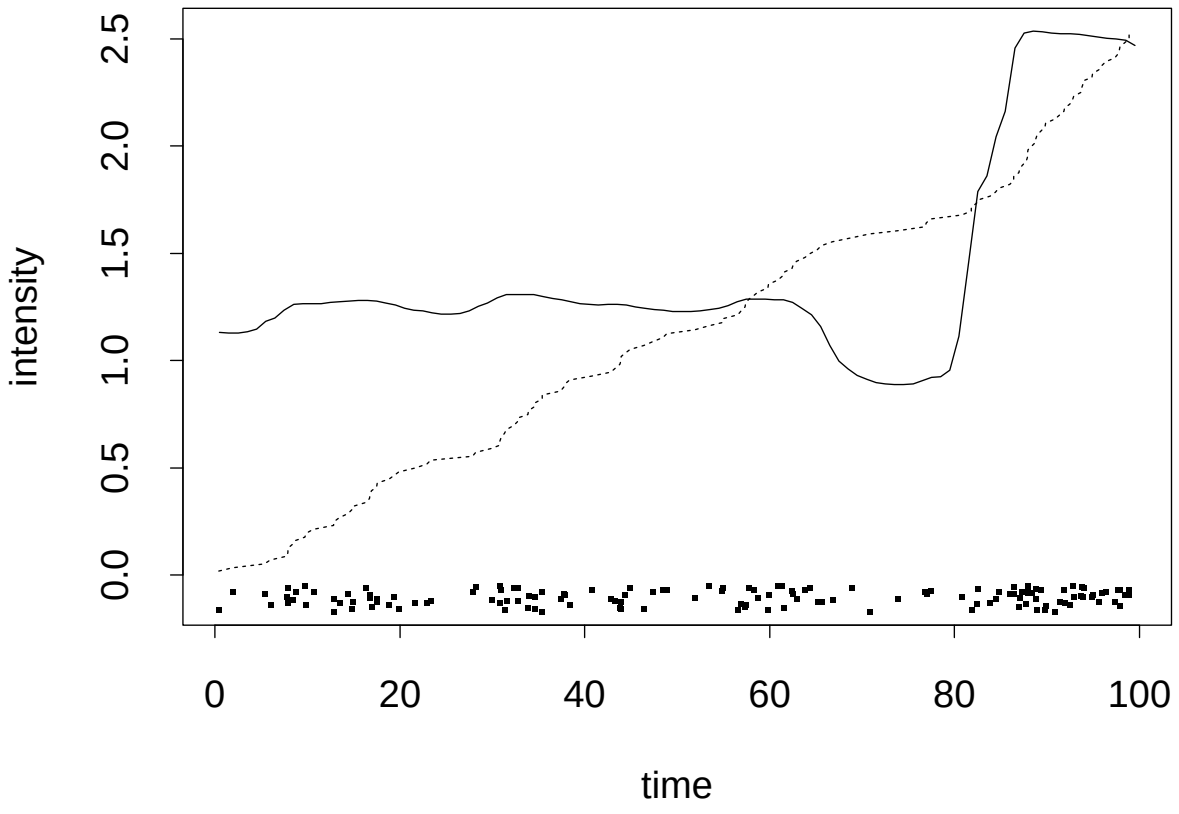


Zero change points is ruled out; $k = 1$ or 2 more probable than under the prior.

Posterior density estimates for change-point positions

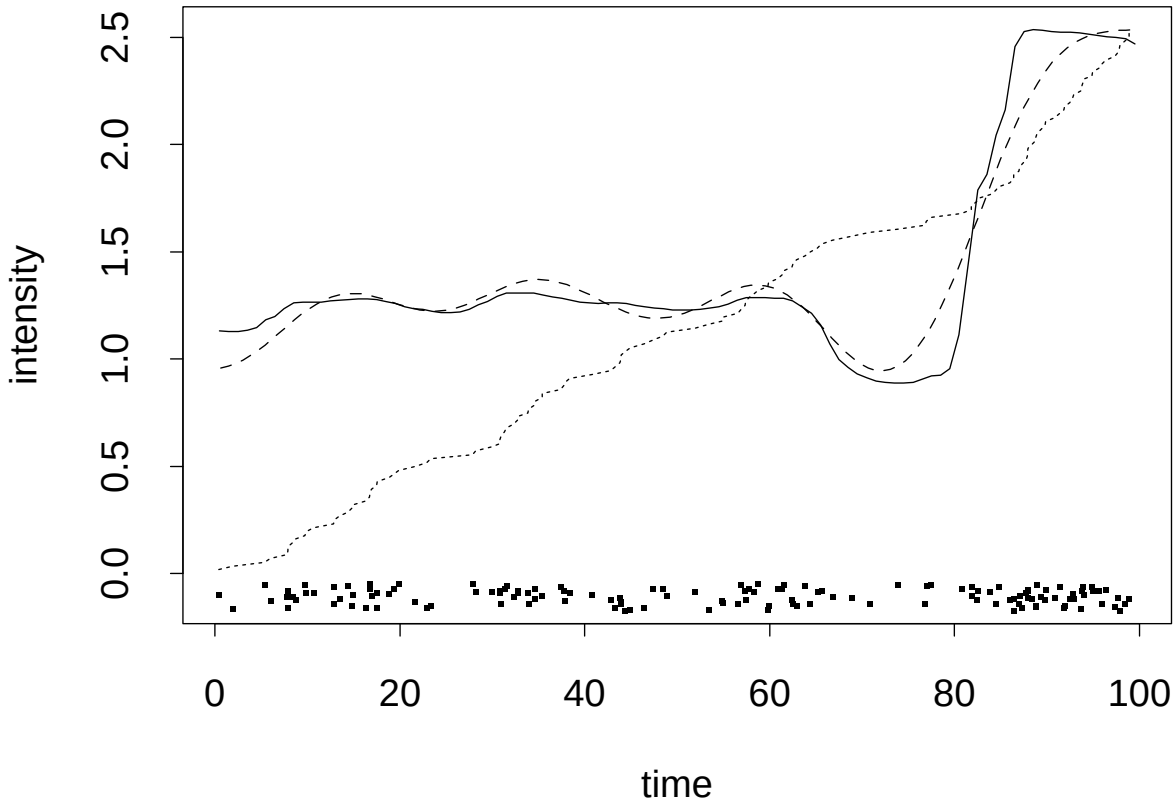


Model-averaged estimate: $E(x(\cdot)|y)$



(the expectation of a random step function is not a step function).

Ordinary smoothing methods (in this case a kernel smoother) can’t match that mean curve



– fixed-bandwidth smoothers either over-smooth the steps, or under-smooth the plateaux.

To follow up

Hjort, N. L. (2002) Topics in nonparametric Bayesian statistics, in *Highly Structured Stochastic Systems*, OUP, to appear. (For details, see `http://www.stats.bris.ac.uk/~peter/L2000/Announce`)

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion). *J. Roy. Statist. Soc. B.*

Green, P. J. and Richardson, S. (2001) Modelling heterogeneity with and without the Dirichlet process, *Scandinavian Journal of Statistics*, **28**, 355–375.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion) *Journal of the Royal Statistical Society*, B, **59**, 731–792.

Green, P. J. and Richardson, S. (2001) Hidden Markov models for disease mapping

Fernández, C. and Green, P. J. (2001) Modelling spatially correlated data via mixtures: a Bayesian approach

Richardson, S., Leblond, L., Jaussent, I. and Green, P. J. (2000) Mixture models in measurement error problems, with reference to epidemiological studies

(the unpublished papers here can be found on the web page below)

My web page:

`http://www.stats.bris.ac.uk/~peter`

My email address:

`P.J.Green@bristol.ac.uk`