Nonparametric Approaches for Heterogeneous Longitudinal Data

Maria De Iorio

Department of Epidemiology and Biostatistics

Imperial College London

June 4th, 2010

Motivation

Random Effects Model

Bayesian Nonparametrics

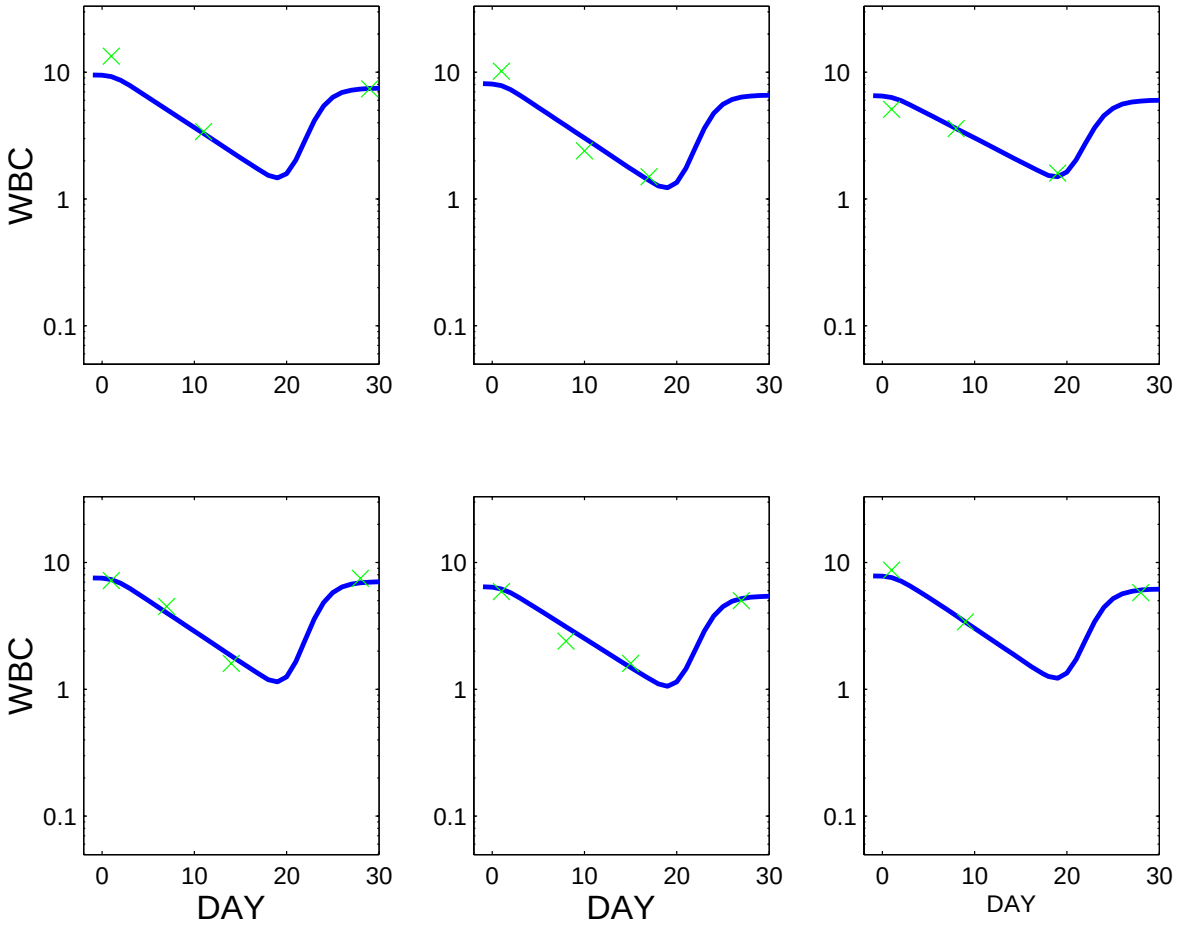
Results and Conclusions

CALGB 8541

The Cancer and Leukemia Group B carried out a large multi-centre randomized study of 3 chemotherapeutic regimens of the drugs cyclophosphamide (CTX), doxorubicin and 5-fluorouracil

- Phase III study: 1572 women were enrolled and randomized.
- 3 treatment arms that contained the same 3 drugs but differed in dose and intensity
- compare the clinical benefits of the 3 regimens for woman with stage II, non-metastatic breast cancer after surgery.
- Focus attention only on the most aggressive regimen (513) women as it causes the most myelosuppression. Analyze WBCs for the first cycle of treatment (28 days).
- Women received the chemotherapy every 4 weeks and WBC measurements were collected only once a week
- there are between 1 and 4 measurements per patient ($\approx$ 3 per patient). Measurements occurred roughly at the same time (days 1, 8, 15, 22)

Typical patients



Problem: Too few data points with which to fit a model able to interpolate with much precision between blood sample times.

→ Introduce information from 2 related earlier phase studies to strengthen inference.

CALGB 8881

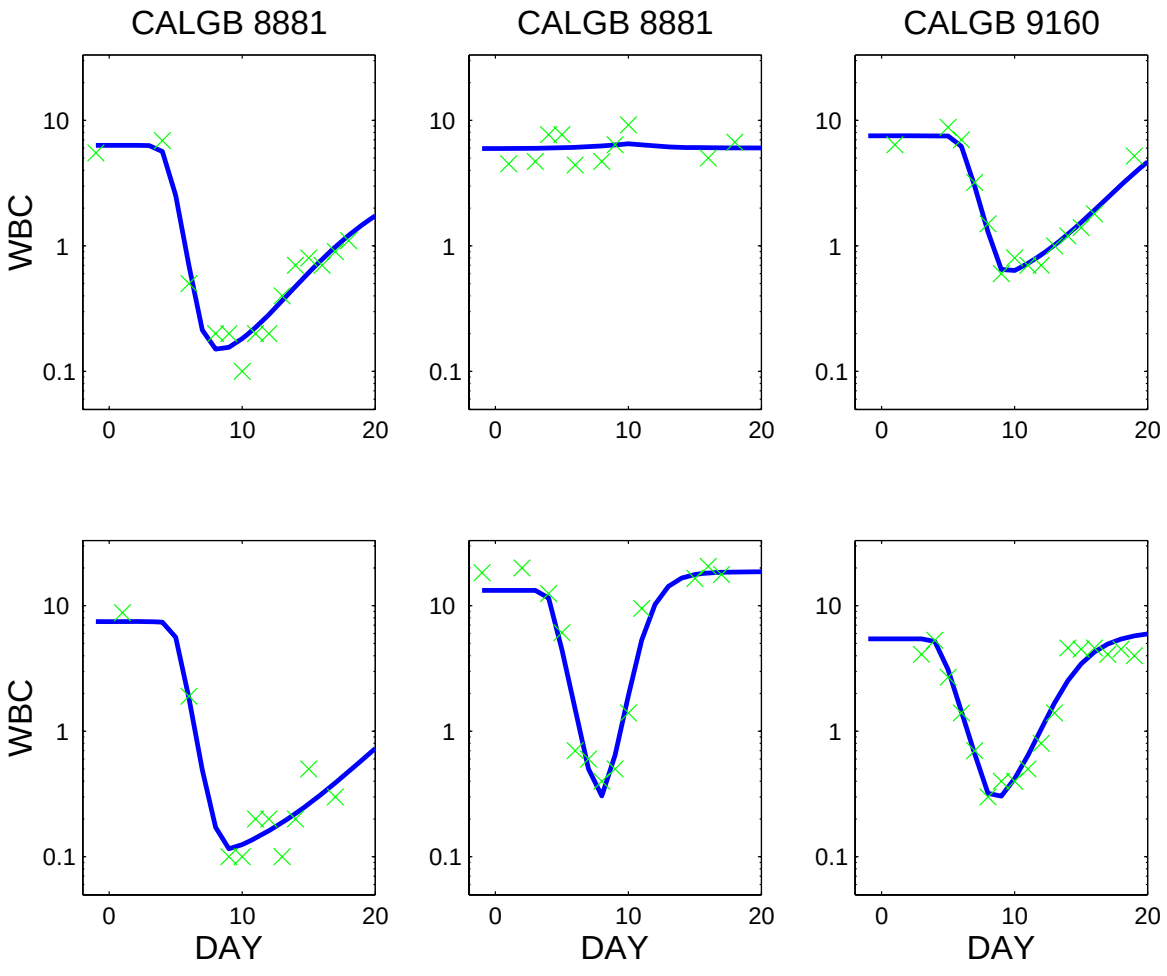
CALGB 8881: Phase I study carried out to determine the highest dose of the anti-cancer agent CTX one can safely deliver every 2 weeks.
CTX causes a drop in WBC. Patients also received GM-CSF (colony stimulating factor given to spur regrowth of blood cells)

- Hematologic toxicity was the primary endpoint
- 46 patients were given different combination of (CTX, GM-CSF) of CTX (grams per square meter of body surface area) and GM-CSF (micrograms per kilogram of body weight): CTX $\in \{1.5, 3.0, 4.5, 6.0\}$ g/m^2 ; GM-CSF $\in \{5.0, 10.0\}$ $\mu g/kg$
- Extensive monitoring: between 4 and 18 measurements per patients ($\approx$ 13 per patient).

CALGB 9160

- built on the experience of CALGB 8881. 46 patients receive CTX = 3 g/m^2 and GM-CSF = 5 $\mu g/kg$
- goal: evaluate the ability of drug amifostine to lessen the toxic effects of relatively high-dose CTX
- patients were randomized to receive amifostine or not.
- there are between 10 and 25 measurements per patients ($\approx$ 15 per patient).

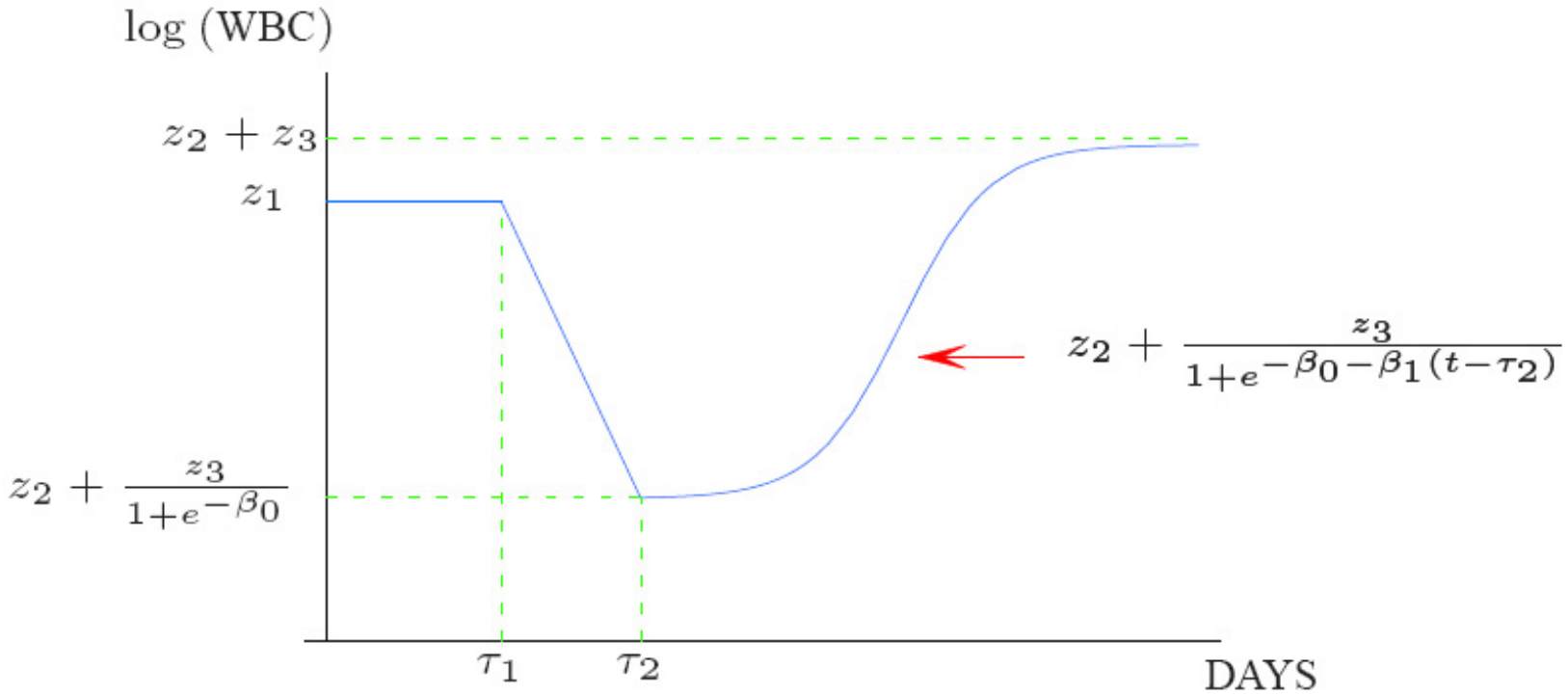
Typical patients from early studies



Note: disparate sampling frequencies for the 3 studies

Nonlinear Regression

Piecewise linear-logistic regression for mean response $y_{ij} = \log(\text{WBC}/1000)$ at time t_{ij}, patient i:



$y_{ij} = f(\boldsymbol{\theta}_i, t_{ij}) + \epsilon_{ij}$

subject-specific random effect vector

$\boldsymbol{\theta}_i = (z_{1i}, z_{2i}, z_{3i}, \tau_{1i}, \tau_{2i}, \beta_{1i})$ and $\beta_0 = -2$

Aim: Meta-Analysis over Related Studies

Goal: combine information in qualitatively different studies to make effective inference.

Want: borrow strength across cancer clinical trials in which the same measurements on WBC counts are collected at different frequencies

Key features of the data: heterogeneous populations and an unbalanced design across the 3 studies of interest

Need: flexible modelling to accommodate heterogeneous population distributions and formalize *borrowing strength* across the studies and across different treatment levels

General Bayesian Model

Top level likelihood:

$$p\left(y_{ij} \mid \theta_i\right)$$

y_{ij} denote the the j-th measurement on the i-th individual, $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$. θ_i denotes denotes a random effects vector for individual i. Typically is a parametric linear/non-linear regression for expected response over time.

Prior Model for the random effect vector:

$$p(\theta_i \mid x_i, \phi)$$

In many application the prior includes a regression on subject-specific covariates

Hyperprior:

$$p(\phi)$$

Random Effects Model

First level of hierarchy: $y_{ij} \mid \theta_i \sim \mathrm{N}\left(f(\theta_i, t_{ij}), \sigma^2\right)$
Second level: random effects distribution $p(\theta_i \mid x_i, \phi)$.

Traditionally $p(\theta_i \mid x_i, \phi)$: Multivariate Normal
Generalization of this approach to account for:

- heterogeneity in the population
- outliers, clustering and over-dispersion
- allow computationally efficient implementation of full posterior inference
- Want: Non-parametric rand. effects dist.'s for θ_i, allowing for dependency on covariate levels.

Dependent Non-parametric Models

Problem: develop dependent nonparametric models for related random probabilities. E.g. the random distribution might be indexed by a categorical covariate indicating the treatment levels in a clinical trial and might represent random effect distribution under the respective treatment combinations.

$$\theta_i \mid x_i \sim p(\theta_i \mid x_i, \phi) = H_{x_i}(\theta_i)$$

- H_x is the random effects distribution for patients with covariates x.
- H_x is a random distribution (or function) $\rightarrow$ non-parametric probability model $p(H_x)$
- Want a dependent prior $p(H_x)$ over $H_x(\cdot)$, covariates $x \in X$.

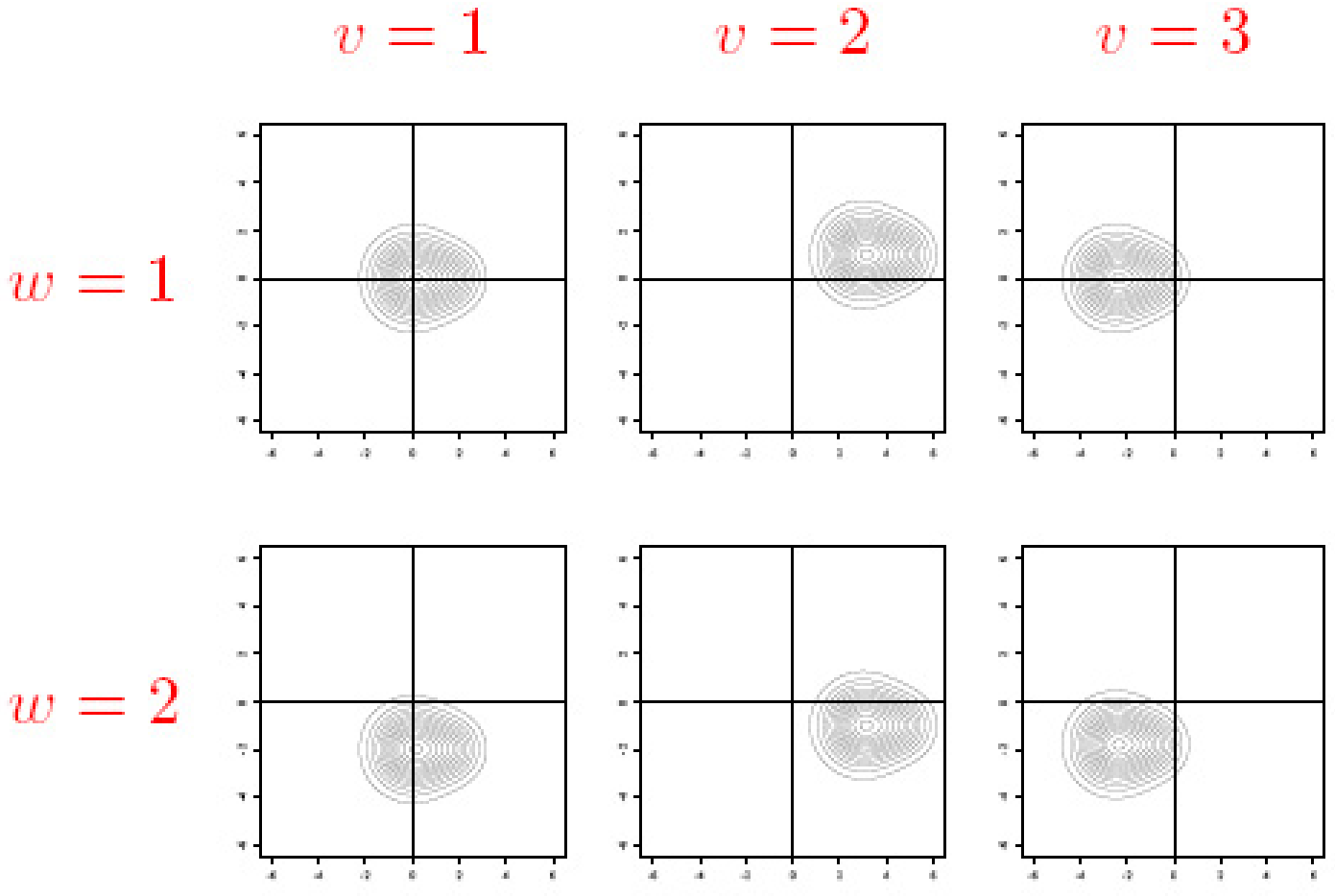
Build hierarchical nonparametric model/prior on data/random effects θ_i

ANOVA for Random Measures/Functions

Array of random distributions $F_x(\cdot)$ for categorical covariates $x = (v, w)$ with

$$v \in \{1, \ldots, V\}, \qquad w \in \{1, \ldots W\}$$

ANOVA of random distributions $F_{vw}(\cdot)$



ANOVA for Random Measures/Functions

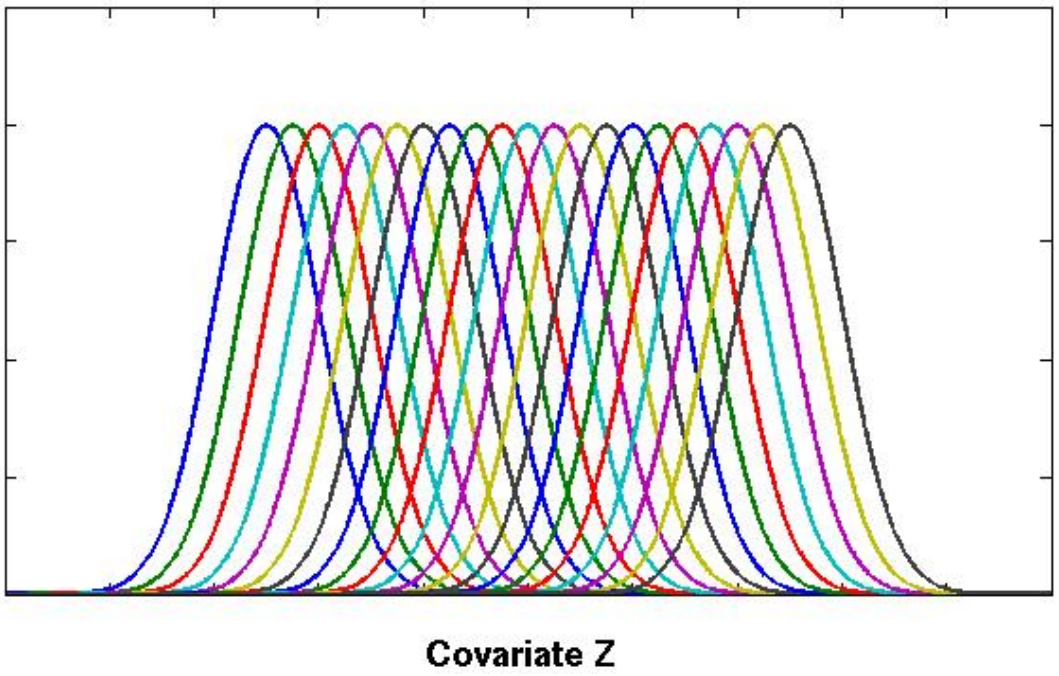
Want:"ANOVA" layout with a different random effect distribution for each combination of covariates

$$
\begin{array}{lll}
x & = & (v, w) \\
H_{x_i} & = & H_{x_j} \quad \text{if} \quad x_i = x_j \\
H_{x_i} & \text{close to} & H_{x_j} \quad \text{if } x_i \text{ and } x_j \text{ only differ in one covariate level} \\
 & \vdots &
\end{array}
$$

Similar idea for continuous covariates

Continuous covariate

Let $z \in Z$ be a continuous covariate, we get a collection of random distribution. The level of dependency is controlled by z.



Dirichlet Process (DP)

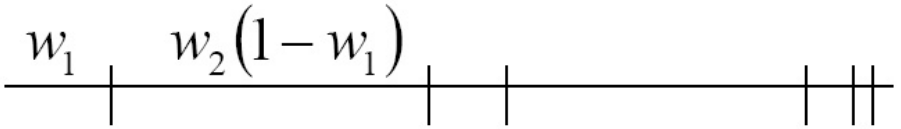
The model is based on the DP (Ferguson 1973)

Probability model on distributions $F \sim DP(M, F^o)$, with measure $F^o = E(F)$ and precision parameter M.

F is a.s. discrete

Sethuraman's stick breaking representation

$$
\begin{aligned}
F &= \sum_{h=1} p_h \delta_{m_h} \\
w_h &\sim \text{Beta}(1, M) \\
p_h &= w_h \prod_{i=1}^{h-1} (1 - w_i), \quad \text{scaled Beta distribution}
\end{aligned}
$$



$$
m_h \overset{iid}{\sim} F^o, \quad h = 1, 2, \ldots
$$

where $\delta(x)$ denotes a point mass at x, p_h are weights of point masses at locations m_h.

G is a discrete distribution, made up of a countably infinite number of point masses. Therefore, there is always a non-zero probability of two observations colliding.

Dirichlet Process Mixtures (DPM)

In many data analysis applications the discreteness is inappropriate.
To remove discreteness: convolution with a continuous kernel

$$\begin{aligned} H(\boldsymbol{\theta}) &= \int p(\boldsymbol{\theta} \mid \mu) \mathrm{d}F(\mu) \\ F &\sim DP(M, F^o) \end{aligned}$$

Dirichlet Process Mixtures (DPM)

or with latent variables μ_i

$$\begin{aligned} F &\sim DP(M, F^o) \\ \mu_i \mid F &\sim F \\ \boldsymbol{\theta} \mid \mu_i &= p(\boldsymbol{\theta} \mid \mu_i) \end{aligned}$$

Nice feature: Mixture is discrete with probability one, and with small M, there can be high probabilities of a finite mixture.

Often $p(\boldsymbol{\theta} \mid \mu) = N(\mu, \sigma^2) \longrightarrow H(\boldsymbol{\theta}) = \sum_{h=1}^{\infty} p_h N(\mu_h, \sigma^2)$

Dependent Dirichlet Process (DDP)

- MacEachern (1999) introduces a probability model for a collection of random distribution $\{F_x, x \in X\}$
- Introduce dependence across x by assuming $m_h = (m_{xh}, x \in X)$ dependent

$$
\begin{aligned}
x = 1: &\quad F_1 = p_1 \delta_{m_{11}} + p_2 \delta_{m_{12}} + \ldots \\
x = 2: &\quad F_2 = p_1 \delta_{m_{21}} + p_2 \delta_{m_{22}} + \ldots \\
x = 3: &\quad F_3 = p_1 \delta_{m_{31}} + p_2 \delta_{m_{32}} + \ldots \\
&\ldots
\end{aligned}
$$

- $m_h = \{m_{xh}, x \in X\} \overset{iid}{\sim} p(m)$, which defines a stochastic process indexed by x, for each fixed h

DDP

- F_x and $F_{x^\star}$ are dependent by virtue of the modelled relationship between the random pairs $\{(m_{xh}, m_{x\star h}) : h = 1, 2, \ldots\}$
- Marginally: $F_x \sim DP(M, F_x^o)$, for all $x \in X$, $m_{xh} \overset{iid}{\sim} F_x^o$
- Computationally easy
- Special case: ANOVA DDP (De Iorio *et al.*, 2004)

ANOVA DDP

- Categorical factors $x = (v, w)$
- Recall $F = \sum p_h \delta_{m_h}$
- Induce dependence across F_x by inducing dependence on point masses
- Introduce dependence across $x = (v, w)$ by assuming an ANOVA model on the locations $\{m_{xh}, x = (v, w), v = 1, \ldots, V, w = 1, \ldots, W\}$

$$m_{xh} = M_h + A_{vh} + B_{wh}$$

with $M_h \sim p_M(M_h), A_{vh} \sim p_{A_v}(A_{vh}), B_{wh} \sim p_{B_w}(B_{wh})$ e.g. $M_h \sim \mathrm{N}(\mu_h, \tau^2)$, etc. and $A_{0h} \equiv B_{0h} \equiv 0$

- Independence across h, dependent - as desired - across x

Interpretation

- Model for the $\{m_{xh}\}$: ordinary ANOVA
- Interpretation M_h: "overall mean"
  A_h, B_h: "main" effects for v and w
- Model is easily generalised to a p-dimensional covariate vector $x = (x_1, \ldots, x_p)$
- Include "interactions", additional factors, inference on contrasts etc. as in ANOVA
- Model allows us to incorporate differential prior information for the various covariate levels
- Easy to include constraints on the estimated effects

Linear DDP

- Extension to continuous covariates (De Iorio *et al* 2009)
- Consider simple case with bivariate covariates $x = (v, z)$ where v is categorical and z is continuous
- Dependence across random distribution by imposing a linear model on the locations (random effects LM)

$$m_{xh} = M_h + A_{vh} + \beta_h z$$

with $M_h \sim p_M(M_h), A_{vh} \sim p_{A_v}(A_{vh})$ and $\beta_h \sim p_\beta(\beta_h)$ and independence across h

- We say $\{F_x : x \in X\} \sim$ Linear DDP(M, p^o)
- The model is easily generalised to more than one continuous covariate

Dirichlet Process Mixture

In many data analysis applications the discreteness is inappropriate.
To remove discreteness: convolution with a continuous kernel

$$\begin{aligned} \theta \mid x, F_x \sim H_x &= \int p(\theta \mid \mu) \mathrm{d}F_x(\mu) \\ \{F_x, x \in X\} &\sim \text{LINEAR DDP}(M, F^o) \end{aligned}$$

where $F_x = \sum_h p_h \delta_{mxh}$, with $m_{xh} \overset{iid}{\sim} F_{0x}$.

Formulation of Linear DDP as DPM

- Consider case with bivariate covariate $x = (v, z)$
- Let $\alpha_h = [M_h, A_{2h}, \ldots, A_{Vh}, \beta_h]$ denote the row vector corresponding to the h-th point mass
- Let d_x denote a design vector such that $\mu_{xh} = \alpha_h d_x$
- Then the linear DDP model can be written as

$$\begin{aligned} p(\theta \mid x, F) &= \int p(\theta \mid \alpha d_x, \Sigma) \mathrm{d}F(\alpha) \\ F &\sim DP(M, F^o) \end{aligned}$$

where $F^o = (p_M, p_A, p_\beta, p_{\sigma^2})$

Large M

- When M is large, F concentrates on F^o, and the model becomes a traditional parametric Bayesian LM
- that is,

$$p(\theta \mid x) = \int p(\theta \mid \alpha d_x, \Sigma) \mathrm{d}F^o(\alpha)$$

- With the additional prior on the ”hyperparameters” of F^o, this is a hierarchical model

Linear DDP as DPM

- For the normal linear model formulation,

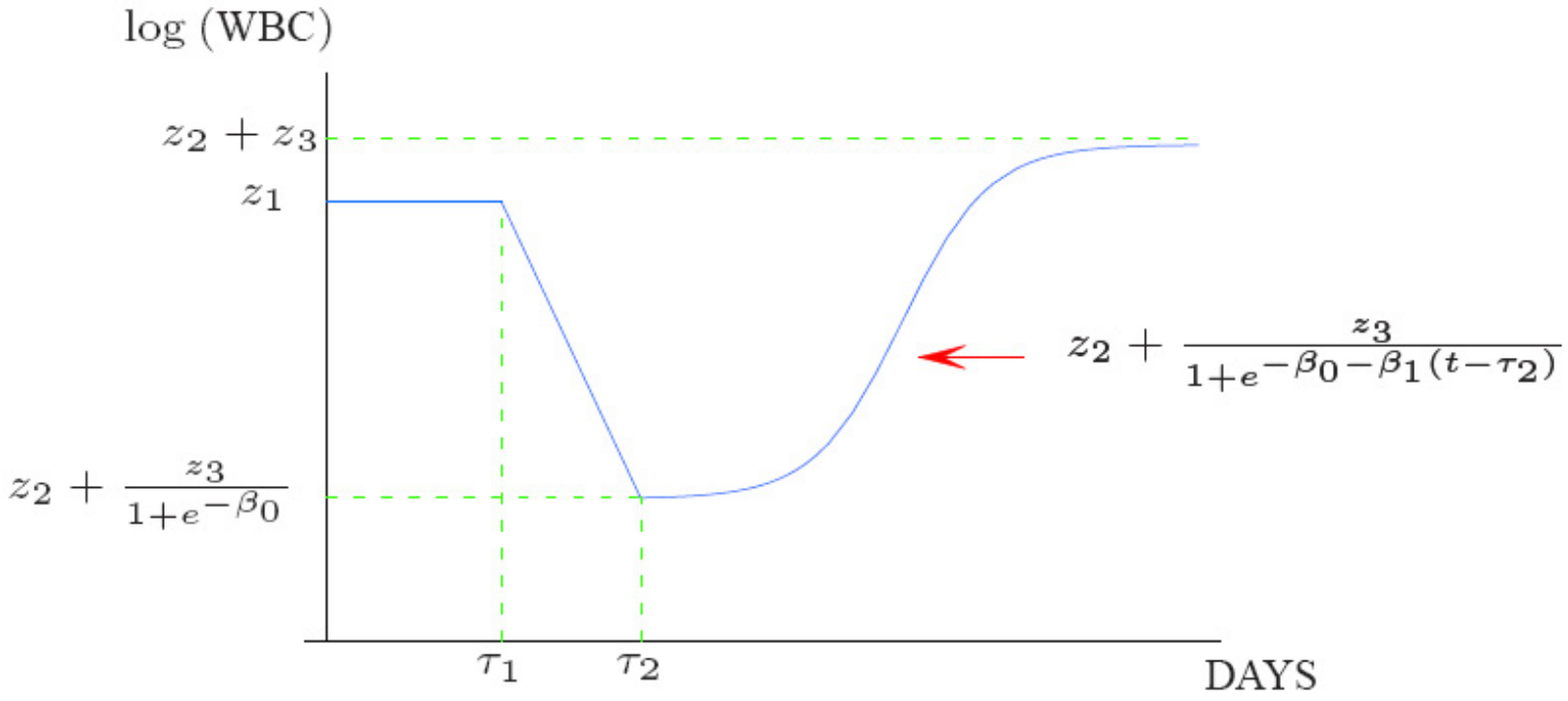
$$E(\boldsymbol{\theta} \mid x, \alpha, F) = m + A_v + \beta z$$

$$\alpha \sim F, \quad F \sim DP(M, F^o)$$

- We are just mixing the linear model using the random mixture F, which for small M will tend to be a finite mixture

Related Longitudinal Studies

Non linear model for mean response $y_{ij} = \log(\text{WBC}/1000)$ at time t_{ij}, patient i:



$yij = f(\theta_i, t_{ij}) + \epsilon_{ij}$

$\theta_i = (z_{1i}, z_{2i}, z_{3i}, \tau_{1i}, \tau_{2i}\beta_{1i})$ and $\beta_0 = -2$

Multiway-ANOVA

ANOVA effects:

- Study $s \in \{8881, 9160, 8541\}$
- CTX $v \in \{1.5, 3.0, 4.5, 6.0\}$
- GM $w \in \{5, 10\}$

Parameters:

$$\mu = [m \mid v_1 \mid v_2 \mid v_3 \mid v_4 \mid w_1 \mid w_2 \mid s_1 \mid s_2 \mid s_3]$$

(5×10) matrix with one column for each ANOVA effect: m corresponds to overall mean, v_1 main effects for CTX = 1.5

Identifiability constraint:

$$s_3 \equiv v_2 \equiv w_1 \equiv 0$$

Hierarchical model

- Dependent prior over measures F_x:

$$\begin{aligned} \{F_x, x \in X\} &\sim \text{LINEAR DDP} \\ F_x &\sim \text{DP}(M, F_x^o) \qquad \text{marginally} \end{aligned}$$

- Convolution w.r.t. Normal kernels (to remove discreteness):

$$H_x = \int \text{N}(\mu, S)\, \text{d}F_x(\mu)$$

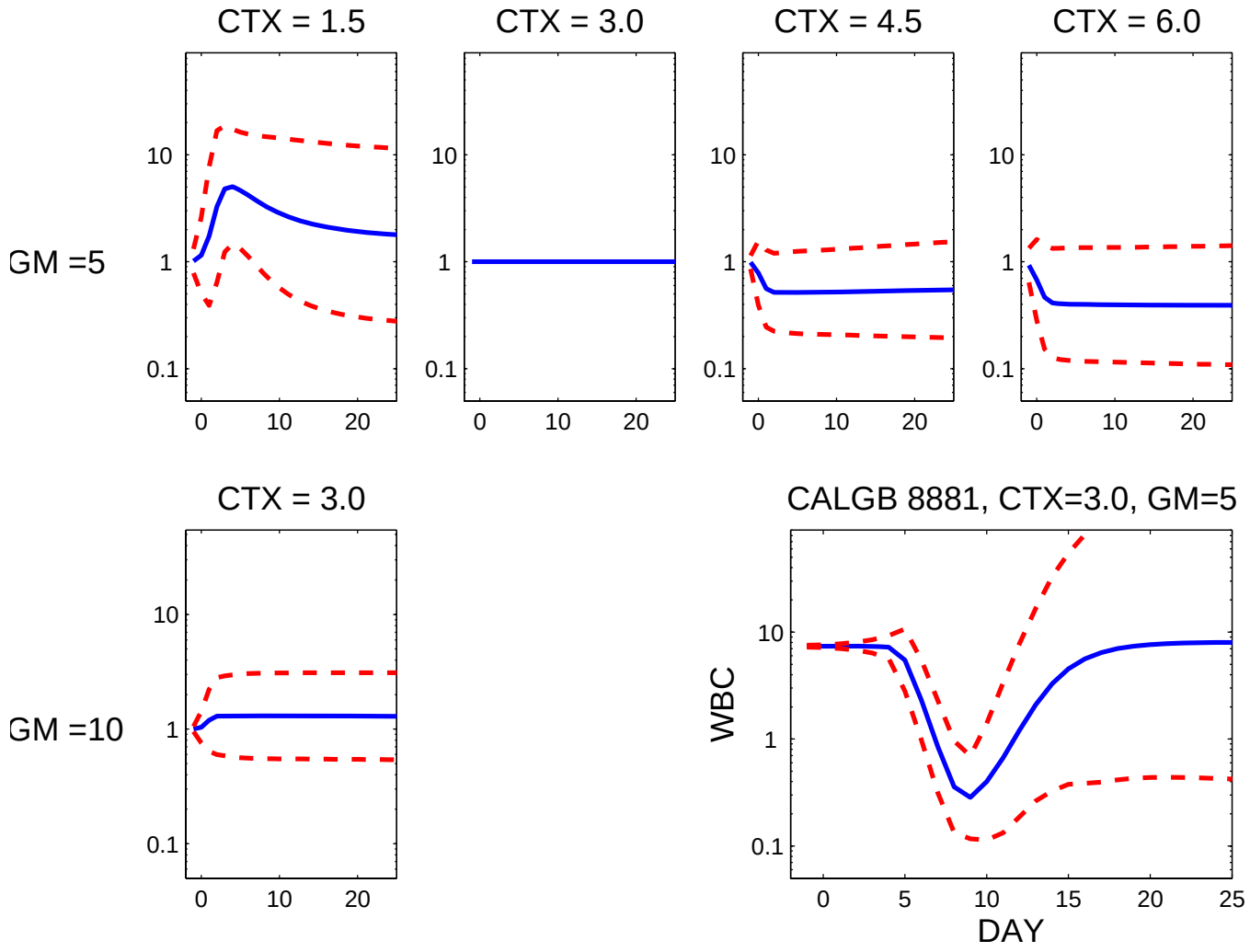
- Random effects vectors:

$$\boldsymbol{\theta}_i \mid x_i = x \sim H_x$$

- Nonlinear regression:

$$y_{ij} = f(\boldsymbol{\theta}_i, t_{ij}) + \epsilon_{ij}$$

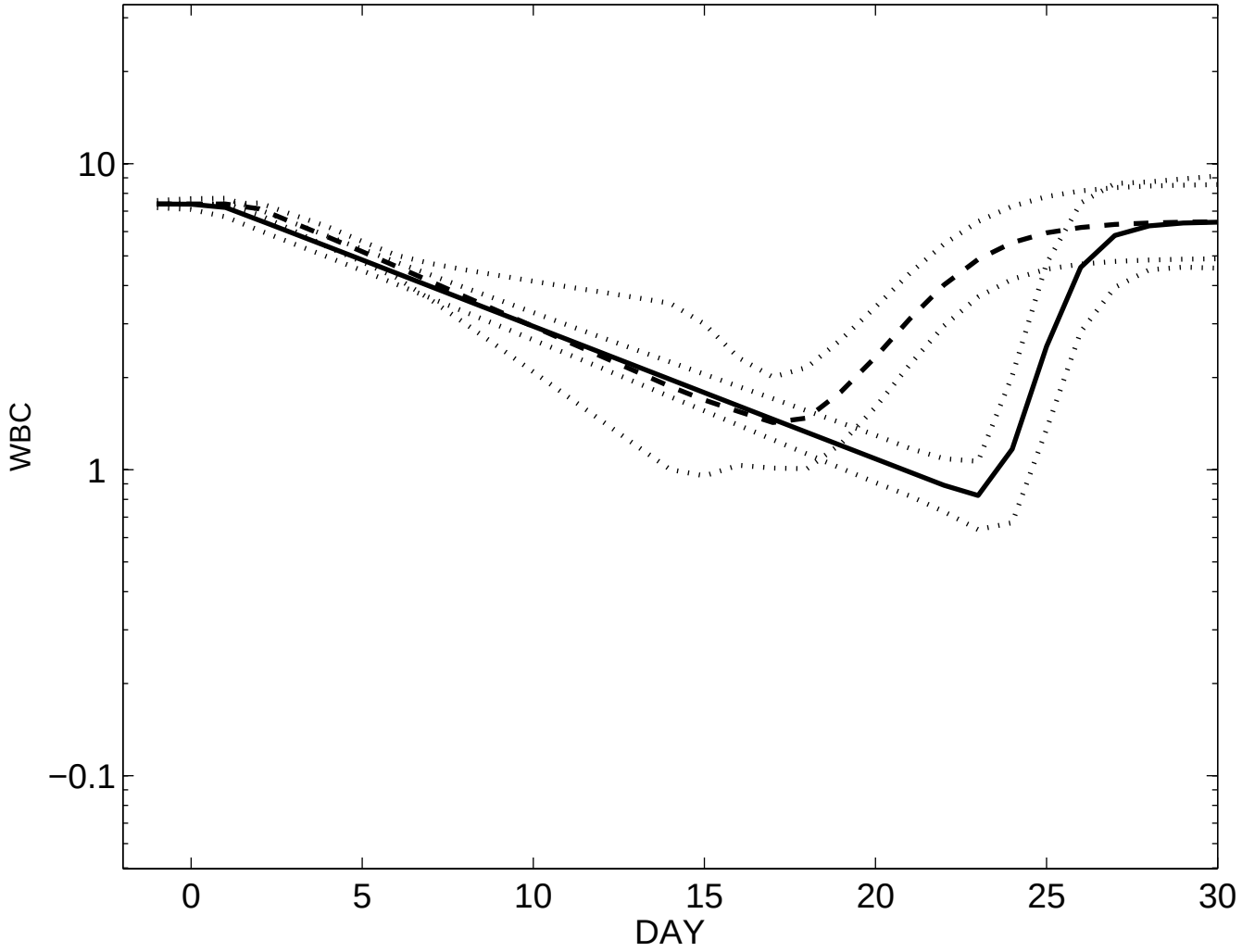
Inference on ANOVA effects



Posterior estimated profiles corresponding to the ANOVA effects of different treatment levels in CALGB 8881. F is high dimensional $\Rightarrow$ posterior inference on the implied nonlinear regression $f(\theta, t)$.

Population Profiles for study 8541

Posterior estimated mean profile for a patient from study 8541: using the hierarchical model (solid) and only the data from study 8541 (dashed). Only CALGB 8541 $\longrightarrow$ more uncertainty about the time of the nadir count and the start of the recovery.



Inference on Myelosuppression

Clinical outcome: Myelosuppression, i.e. a profound lowering of a person's bone marrow activity leading to a reduction in the number of platelets, red blood cells and white blood cells. Common side effect of anticancer drug therapy.

Consequences on inference about the extent of myelosuppression (e. g. nadir count, number of days the patient's WBC are below some threshold value).

Number of days that the mean WBC is below the critical value of WBC = 1000

Hierarchical model: posterior mean = 5.15

Only CALGB 8541: posterior mean = 1.04

Huge difference due to the fact that relatively few observations under study CALGB 8541 do not allow precise information about the day of recovery.

Conclusions

- We have introduced a probability model for dependent random distributions
- ease of interpretation
- facility to impose structure
- we have exploited the model to define inference across related, non-exchangeable studies
- extension to a variety of contexts in which the data are collected at different resolutions by design (e.g. drug development)
- efficient computation (R packages available)
- MCMC scheme relies on the conjugacy of the base measure and mixing kernel (MacEachern and Müller 1998; Neal 2000; Griffin and Walker 2009)

Acknowledgements

Peter Müller

Gary Rosner

Steve MacEachern

Wesley Johnson