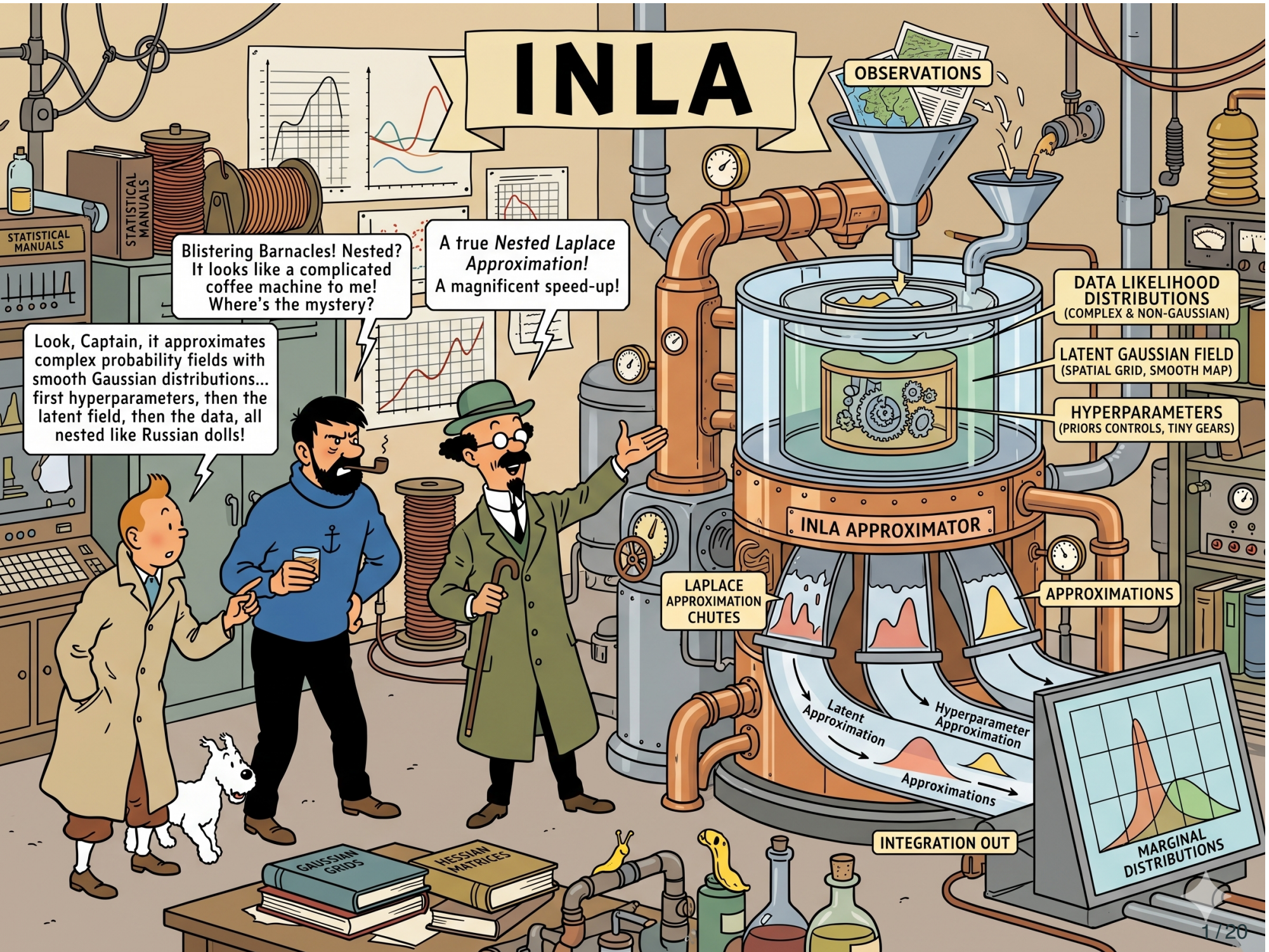

INLA
OBSERVATIONS
STATISTICAL MANUALS
STATISTICAL MANUALS
Blistering Barnacles! Nested? It looks like a complicated coffee machine to me! Where's the mystery?
A true *Nested Laplace Approximation!* A magnificent speed-up!
Look, Captain, it approximates complex probability fields with smooth Gaussian distributions... first hyperparameters, then the latent field, then the data, all nested like Russian dolls!
DATA LIKELIHOOD DISTRIBUTIONS (COMPLEX & NON-GAUSSIAN)
LATENT GAUSSIAN FIELD (SPATIAL GRID, SMOOTH MAP)
HYPERPARAMETERS (PRIORS CONTROLS, TINY GEARS)
INLA APPROXIMATOR
LAPLACE APPROXIMATION CHUTES
APPROXIMATIONS
Latent Approximation
Hyperparameter Approximation
Approximations
INTEGRATION OUT
MARGINAL DISTRIBUTIONS
GAUSSIAN GRIDS
HESSIAN MATRICES

INLA

- The big breakthrough of Bayesian inference has been largely driven by simulation-based methods introduced in 1900's (i.e., MCMC).
- But deterministic approximation to the full Bayesian inference has many practical advantages: fast, easier for applied users, easier to automate, ...
- The EM algorithm we looked at in the previous lecture is one such example, but it only estimates the MAP.
- We now turn to a different approach that approximates posterior marginal distributions of parameters: the integrated nested Laplace approximation (INLA).
- Unlike Stan, INLA is especially designed for latent Gaussian models, using a deterministic algorithm based on Laplace approximation.
- A good reference book for INLA is: Blangiardo, Marta, and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

INLA: the class of models

- Consider models in the following form. For $y = (y_1, ..., y_n)$

$$y_i \sim p(\eta_i)$$
$$\eta_i = \beta_0 + \sum_m \beta_m x_{mi} + \sum_\ell f_\ell(z_{\ell i})$$

  where $p()$ specifies the distribution of y parameterized by η_i, and η_i can be decomposed into additive terms including the intercept, linear effects from covariates x and random effects f_ℓ.
- We will not go into details of random effects for now, but think about f_ℓ as smoothing terms that can include correlated residuals (e.g., over space or time).

INLA: the class of models

- Denote all parameters in the additive model as $\theta = (\beta_0, \beta, f)$ and all hyperparameters $\psi = (\psi_1, ..., \psi_K)$. For simpler presentation, assume each y_i is connected with only one θ_i, then

$$p(y|\theta, \psi) = \prod_i p(y_i|\theta_i, \psi)$$

  This assumption can be relaxed with y_i depend on a linear combination of θ_i's.

- Key assumption 1: multivariate normal prior on θ with mean 0 and precision matrix $Q(\psi)$, the posterior is $p(\theta, \psi|y) \propto p(\psi)p(\theta|\psi)p(y|\theta, \psi)$
- Key assumption 2: the log likelihood function $\log p(y_i|\theta_i, \psi)$ is log concave in terms of θ. This is satisfied by most of the standard likelihood functions we typically use.

Example

$$y_{it} \sim \text{Poisson}(E_{it}\rho_{it})$$

$$\log(\rho_{it}) = \beta_0 + u_i + v_i + \gamma_t$$

$$u_j|u_{-j} \sim N(\frac{1}{N_i}\sum_{j\in ne(i)} u_j, \sigma_u^2)$$

$$v_i \sim N(0, \sigma_v^2)$$

$$\gamma_t|\gamma_{t-1} \sim N(\gamma_{t-1}, \sigma_\gamma^2)$$

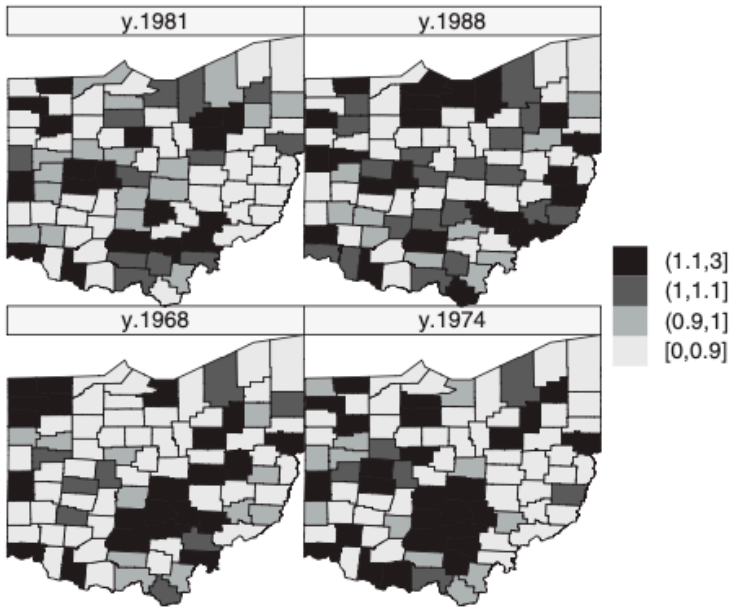


Figure 1.12 Distribution of standardized mortality rates of lung cancer in 88 counties in Ohio (USA) during 1968, 1974, 1981, and 1988.

INLA: step by step

- Suppose we are interested in knowing the marginal distributions of θ and ψ

$$p(\theta_i|y) = \int p(\theta_i|\psi, y)p(\psi|y)d\psi$$

$$p(\psi_k|y) = \int p(\psi|y)d\psi_{-k}$$

- We need to compute two quantities, $p(\psi|y)$ and $p(\theta_i|\psi, y)$.
- For the former, we note it holds for any value of θ,

$$\begin{aligned} p(\psi|y) &= \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)} \\ &\propto \frac{p(y|\theta, \psi)p(\theta|\psi)p(\psi)}{p(\theta|\psi, y)} \\ &\approx \frac{p(y|\theta, \psi)p(\theta|\psi)p(\psi)}{\tilde{p}(\theta|\psi, y)}\Big|_{\theta=\theta^\star(\psi)} \end{aligned}$$

  where $\tilde{p}(\theta|\psi, y)$ is the Laplace approximation of $p(\theta|\psi, y)$ and $\theta^\star(\psi)$ is its mode.
- That is, if we can numerically approximate the normalizing constant, we can directly get posterior marginal distribution!

Laplace approximation

- Laplace approximation is a pre-MCMC era technique used to approximate integrals.
- Consider the estimation of

$$\int \exp(nf(x))dx$$

  where $nf(x)$ can be thought of the log likelihood function.
- It can be approximated by matching the target function by a Gaussian. Consider a Taylor expansion at $x_0 = \arg\max f(x)$,

$$\int \exp\left(n(f(x_0) + \frac{1}{2}(x - x_0)^2 f''(x))\right)dx = \exp(nf(x_0))\sqrt{\frac{2\pi}{-nf''(x_0)}}$$

- The approximation is exact as $n \to \infty$. When n is finite, the approximation is better when $f(x)$ is close to log Gaussian density.

Why approximating $p(\theta|\psi, y)$?

- Gaussian approximation here is accurate as $p(\theta|\psi, y)$ is usually close to Gaussian, because it is a priori Gaussian.
- And y generally is not very informative. To see this,

$$p(\theta|\psi, y) \propto \exp(-\frac{1}{2}\theta^T Q(\psi)\theta + \sum_i \log p(y_i|\theta_i, \psi))$$

  which does not have product terms $\theta_i\theta_j$ so all the impact from conditioning on the observations is only on the diagonal. The log concavity of the likelihood also ensures the sum is close to another log Gaussian term.
- The dimension of ϕ is usually low, e.g., three variance terms (log precision to be more exact) in the Ohio example.
- Note: if we are to do MCMC, we may sample (θ, ψ) from the posterior using the Laplace approximation:
  1. Draw ψ from the approximated $p(\psi|y)$ on a grid of ψ values.
  2. Draw θ from $\tilde{p}(\theta|\psi, y)$
  3. Accept or reject (ψ, θ) jointly

Compare Laplace approximation and Gaussian approximation

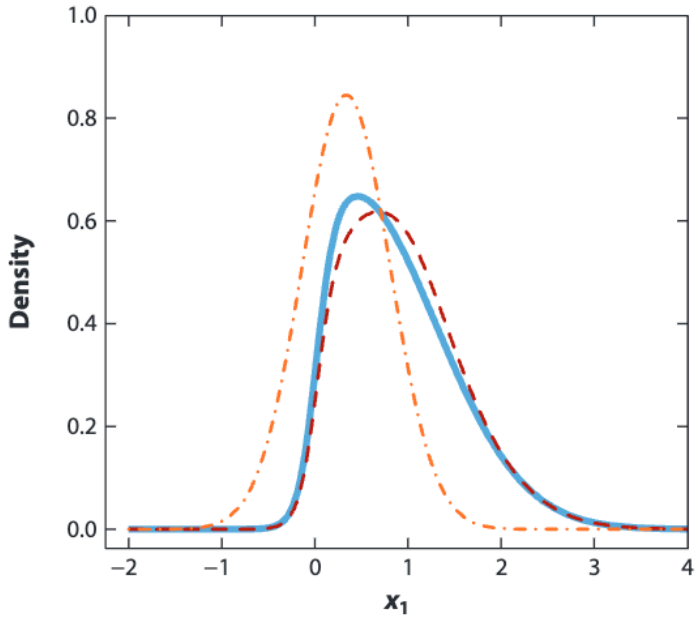


Figure 1

The true marginal (*solid blue line*), the Laplace approximation (*dashed red line*) and the Gaussian approximation (*dot-dashed orange line*).

Let us now discuss a simplistic, but realistic, model in two dimensions $\boldsymbol{x} = (x_1, x_2)^T$, where

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \boldsymbol{x}\right) \prod_{i=1}^{2} \frac{\exp(c\, x_i)}{1 + \exp(c\, x_i)} \tag{8}$$

for a constant $c > 0$ and $\rho \geq 0$. This is the same functional form as we get from two Bernoulli successes, using a logit-link. Using the constant c is an alternative to scaling the Gaussian part, and the case where $\rho < 0$ is similar. The task now is to approximate $\pi(x_1) = \pi(x_1, x_2)/\pi(x_2|x_1)$,

Approximating $p(\theta_i|\psi, y)$?

- $p(\theta_i|\psi, y)$ is trickier.
- $\boldsymbol{\theta}$ is assumed to be normal a priori, but taking the marginal distribution of direct Gaussian approximation of $p(\theta|\psi, y)$ is not accurate enough.
- We can use Laplace approximation marginally for each θ_i,

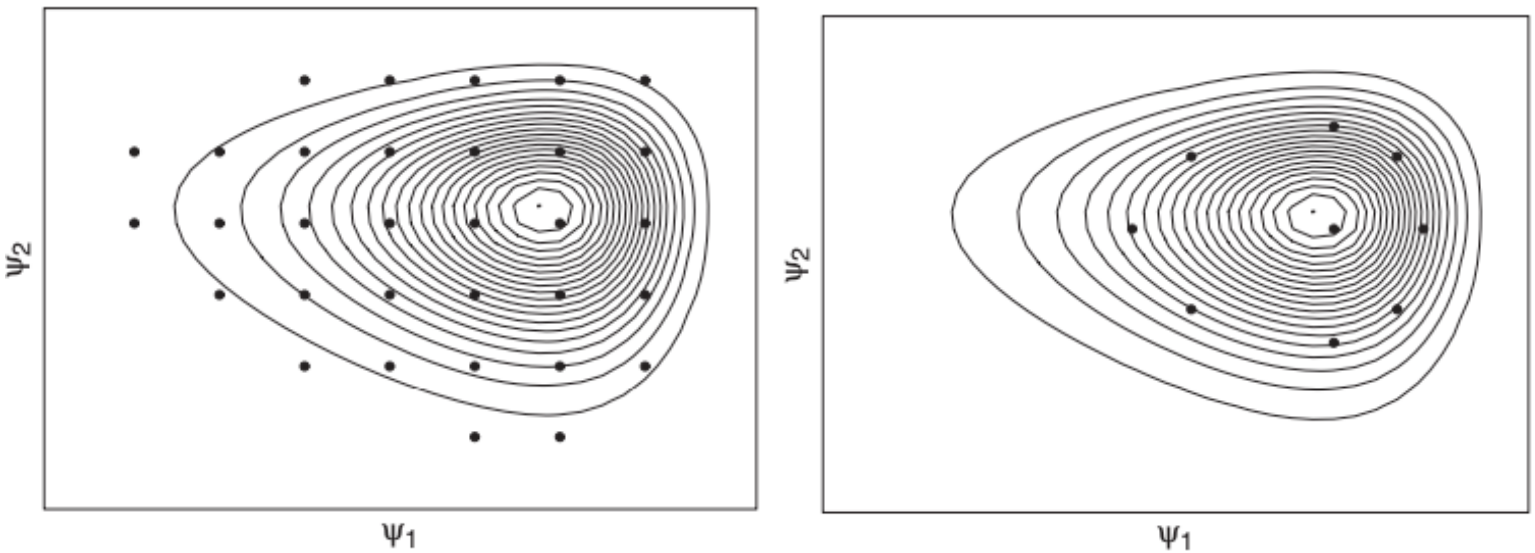
$$p(\theta_i|\psi, y) \approx \frac{p(\theta, \psi|y)}{\tilde{p}(\theta_{-i}|\theta_i, \psi, y)}|_{\theta_{-i}=\theta^{\star}_{-i}(\theta_i,\psi)}$$

  but it is computationally expensive to recompute the precision matrix every time (as it depends on θ_{-i}), especially when θ is high dimensional.
- Instead INLA defaults to use a method named 'simplified Laplace approximation' (Rue, Martino, and Chopin (2009)) that uses the third order Taylor expansion to approximate $p(\theta_i|\psi, y)$ with an extra spline correction

$$p(\theta_i|\psi, y) \approx N(\theta_i; m_i(\boldsymbol{\theta}), V_i(\boldsymbol{\theta})) \exp(spline(\theta_i))$$

Step by step (roughly)

- Explore the joint distribution of $\tilde{p}(\phi|y)$ and find a grid of good integration points associated with the bulk of the mass and associated area weights Δ.



- For each integration point $\psi^\star$ and parameter θ_i, evaluate the approximated marginal $\tilde{p}(\theta_i|\psi^\star, y)$ for selected values of θ_i.
- For each i obtain the marginal posterior using numerical integration

$$\tilde{p}(\theta_i|y) \approx \sum_{\psi^\star} \tilde{p}(\theta_i|\psi^\star, y)\tilde{p}(\psi^\star|y)\Delta^\star$$

The INLA package in R

- There are many more numerical optimizations going on behind the scenes, but the good news is that there is a highly efficient implementation of the INLA procedure in R.
- Similar to Stan, here is just a fairly brief preview of the INLA package.

```
install.packages("INLA",repos=c(getOption("repos"),
        INLA="https://inla.r-inla-download.org/R/stable"),
        dep=TRUE)
library(INLA)
```

- The INLA package also has been extended in several different ways including drawing posterior samples from the posterior joint distribution and dealing with models that deviate from latent Gaussian models slightly.

Relationship between temperature and PM10

- We consider daily PM10 concentration and temperature in Salt Lake City between 1987 and 2000.
- We want to assess the presence of a linear relationship between temperature and air pollution concentration.
- A simple linear model is

$$PM10_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 \text{Temp}_i$$

Relationship between temperature and PM10

Here we specifies priors for the fixed effect $\beta_1 \sim N(0, 10000)$, intercept $\beta_0 \sim N(0, 1)$, and precision $1/\sigma^2 \sim Gamma(1, 0.00005)$, just for the demonstration.

```
data <- read.csv("https://sites.google.com/a/r-inla.org/stbook/NMMAPSraw.csv")
formula <-  pm10 ~ 1 + temperature
model.linear <- inla(formula, family = "gaussian", data = data,
                control.fixed=list(mean=0, prec=1,
                                   mean.intercept=0, prec.intercept=0.0001),
                control.family=list(hyper=list(prec=
                                    list(prior="loggamma",param=c(1, 5e-5)))))
```

Relationship between temperature and PM10

```
> round(model.linear$summary.fixed, 3)
              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept) 38.725 0.983     36.794   38.725     40.653 38.725   0
temperature -0.094 0.017     -0.128   -0.094     -0.060 -0.094   0
> summary(lm(formula, data = data))

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-50.584 -14.941  -4.029   9.222 273.208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.72517    0.98373  39.366  < 2e-16 ***
temperature -0.09400    0.01746  -5.384 7.62e-08 ***
```

Relationship between temperature and PM10

The results from the linear regression seems counter-intuitive. Higher temperature is associated with decreased PM10. If we plot the data, we can clearly see the cyclic pattern in the PM10 measurements. A more realistic approach is to take into account the temporal correlation. Let us consider a simple case here by including another random effect in the form of first order random walk:

$$PM10_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 \text{Temp}_i + u_i$$
$$u_i | u_{i-1} \sim N(u_{i-1}, \sigma_u^2)$$

```
formula <- pm10 ~  1 +  temperature + f(id, model="rw1")
model.rw1 <- inla(formula, family="gaussian", data = data)
```

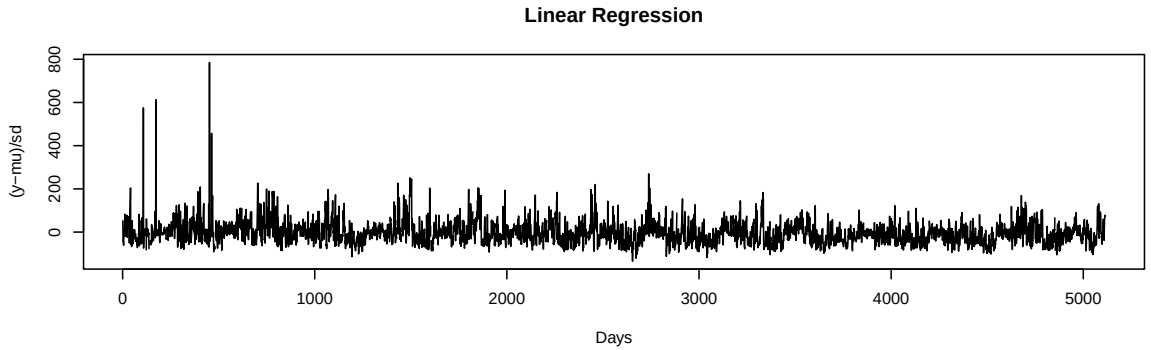
Relationship between temperature and PM10

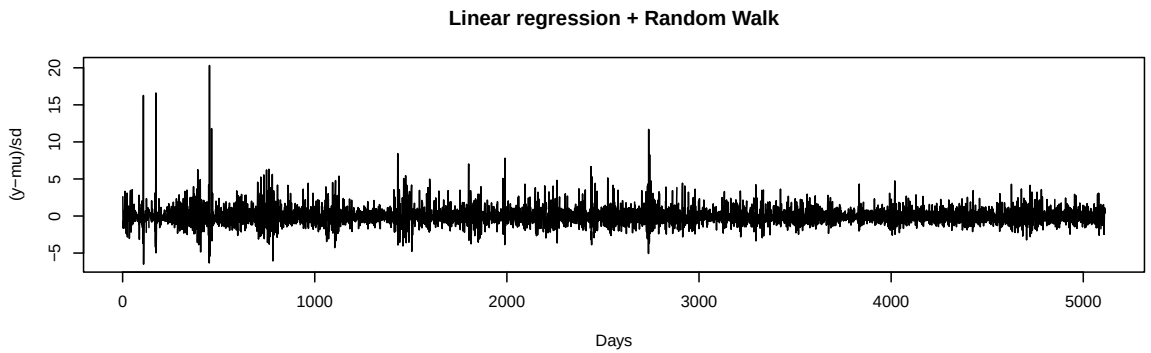
We can compare the standardized residuals $(y_i - \hat{\mu}_i)/sd(\hat{\mu}_i)$ from the two models

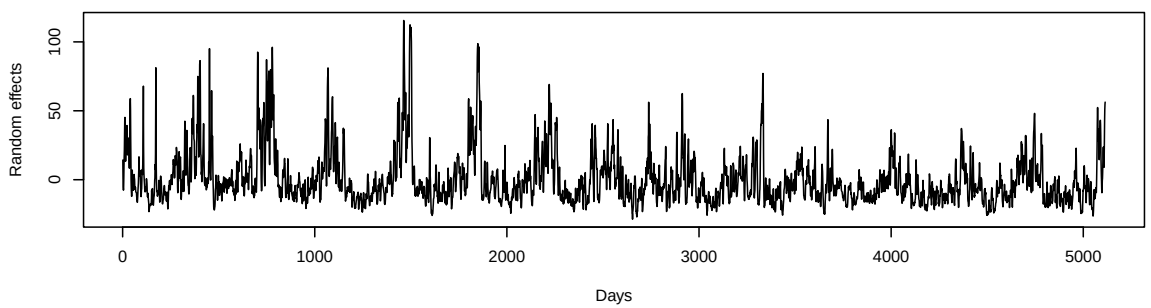
```
res.linear <- (data$pm10 - model.linear$summary.fitted.values$mean) /
              (model.linear$summary.fitted.values$sd)
res.rw1 <- (data$pm10 - model.rw1$summary.fitted.values$mean) /
              (model.rw1$summary.fitted.values$sd)

plot(res.linear, xlab="Days", ylab = "(y-mu)/sd", type = 'l',
             main = "Linear Regression")
plot(res.rw1, xlab="Days", ylab = "(y-mu)/sd", type = 'l',
             main = "Linear regression + Random Walk")
plot(model.rw1$summary.random$id$mean, xlab = "Days",
             ylab = "Random effects", type = 'l')
```

Relationship between temperature and PM10







Relationship between temperature and PM10

Running the model now shows the PM10 increases as temperatures increase.

```
plot(model.linear$marginals.fixed[[1]], type = "l", xlab = "beta0",
      xlim = c(0, 45))
plot(model.linear$marginals.fixed[[2]], type = "l", xlab = "beta1",
      xlim = c(-0.3, 0.8))
plot(model.rw1$marginals.fixed[[1]], type = "l", xlab = "beta0",
      xlim = c(0, 45))
plot(model.rw1$marginals.fixed[[2]], type = "l", xlab = "beta1",
      xlim = c(-0.3, 0.8))
```

Relationship between temperature and PM10