# Chapter 8: Modelling accounting for data collection

Highly recommended to read. Very informative, but also dense chapter.

- We need to model the data collection unless it is ignorable
- We need to know when data collection is ignorable

# Chapter 8: Modelling accounting for data collection

Highly recommended to read. Very informative, but also dense chapter.

- We need to model the data collection unless it is ignorable
- We need to know when data collection is ignorable
- Data collection
  - Sample surveys
  - Designed experiments
  - Randomization
  - Observational studies
  - Censoring and truncation

# Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$
    we can model just $p(y|x, \theta)$

# Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$
    we can model just $p(y|x, \theta)$
- Gaussian linear model with conjugate prior
  - the conditional posterior is multivariate normal
  - with fixed prior on weights, the joint posterior is N-Inv-$\chi^2$

## Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$
    we can model just $p(y|x, \theta)$
- Gaussian linear model with conjugate prior
  - the conditional posterior is multivariate normal
  - with fixed prior on weights, the joint posterior is N-Inv-$\chi^2$
  - these properties are sometimes useful and thus good to know, but with probabilistic programming less often needed

# Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$ we can model just $p(y|x, \theta)$
- Gaussian linear model with conjugate prior
  - the conditional posterior is multivariate normal
  - with fixed prior on weights, the joint posterior is N-Inv-$\chi^2$
  - these properties are sometimes useful and thus good to know, but with probabilistic programming less often needed
- Bit on causal analysis (see much more in ROS)

# Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$
    we can model just $p(y|x, \theta)$
- Gaussian linear model with conjugate prior
  - the conditional posterior is multivariate normal
  - with fixed prior on weights, the joint posterior is N-Inv-$\chi^2$
  - these properties are sometimes useful and thus good to know, but with probabilistic programming less often needed
- Bit on causal analysis (see much more in ROS)
- Assembling matrix of explanatory variables
  - identifiability, collinearity, nonlinear relations, indicator and categorical variables, interactions
  - variable selection is not much discussed (see lectures 9.2,9.3)

# Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$ we can model just $p(y|x, \theta)$
- Gaussian linear model with conjugate prior
  - the conditional posterior is multivariate normal
  - with fixed prior on weights, the joint posterior is N-Inv-$\chi^2$
  - these properties are sometimes useful and thus good to know, but with probabilistic programming less often needed
- Bit on causal analysis (see much more in ROS)
- Assembling matrix of explanatory variables
  - identifiability, collinearity, nonlinear relations, indicator and categorical variables, interactions
  - variable selection is not much discussed (see lectures 9.2,9.3)
- Regularization
  - not much discussed (see more in lecture 9.3 and e.g. https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html)

# Chapter 14: Introduction to regression models

- Justification of conditional modeling
  - if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$ we can model just $p(y|x, \theta)$
- Gaussian linear model with conjugate prior
  - the conditional posterior is multivariate normal
  - with fixed prior on weights, the joint posterior is N-Inv-$\chi^2$
  - these properties are sometimes useful and thus good to know, but with probabilistic programming less often needed
- Bit on causal analysis (see much more in ROS)
- Assembling matrix of explanatory variables
  - identifiability, collinearity, nonlinear relations, indicator and categorical variables, interactions
  - variable selection is not much discussed (see lectures 9.2,9.3)
- Regularization
  - not much discussed (see more in lecture 9.3 and e.g. https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html)
- Unequal variances and correlations

## Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted

## Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
  - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori

## Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
  - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
  - when the amount of penalty is increased, marginal modes of weak effects go to zero first

# Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
  - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
  - when the amount of penalty is increased, marginal modes of weak effects go to zero first
  - when the amount of penalty is increased, also the relevant coefficients are shrunk towards zero

# Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
  - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
  - when the amount of penalty is increased, marginal modes of weak effects go to zero first
  - when the amount of penalty is increased, also the relevant coefficients are shrunk towards zero
  - sometimes relaxed lasso is used, where after variable selection coefficients are re-estimated

## Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
    - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
    - when the amount of penalty is increased, marginal modes of weak effects go to zero first
    - when the amount of penalty is increased, also the relevant coefficients are shrunk towards zero
    - sometimes relaxed lasso is used, where after variable selection coefficients are re-estimated
- Bayesian lasso uses Laplace distribution as prior
    - Laplace prior is equivalent to L1 penalty

## Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
    - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
    - when the amount of penalty is increased, marginal modes of weak effects go to zero first
    - when the amount of penalty is increased, also the relevant coefficients are shrunk towards zero
    - sometimes relaxed lasso is used, where after variable selection coefficients are re-estimated
- Bayesian lasso uses Laplace distribution as prior
    - Laplace prior is equivalent to L1 penalty
    - but the Bayesian inference includes distribution for parameters and that distribution doesn't shrink to a point at zero, even if the mode would be at zero

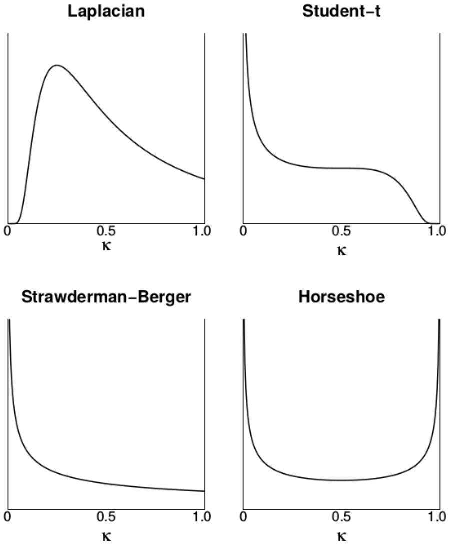
## Lasso and Bayesian lasso

- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
  - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
  - when the amount of penalty is increased, marginal modes of weak effects go to zero first
  - when the amount of penalty is increased, also the relevant coefficients are shrunk towards zero
  - sometimes relaxed lasso is used, where after variable selection coefficients are re-estimated
- Bayesian lasso uses Laplace distribution as prior
  - Laplace prior is equivalent to L1 penalty
  - but the Bayesian inference includes distribution for parameters and that distribution doesn't shrink to a point at zero, even if the mode would be at zero
  - empirically better results obtained with more sparse priors

## Lasso and Bayesian lasso

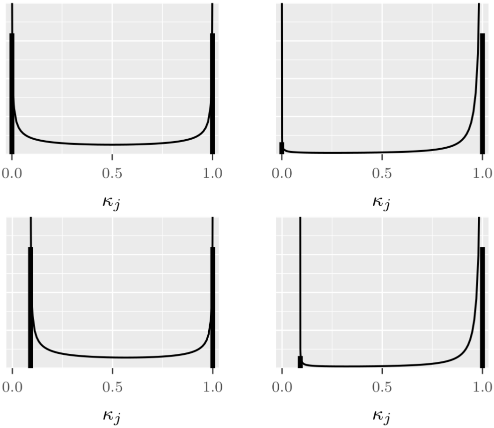
- Lasso is penalized maximum likelihood linear regression, with L1 one penalty where the amount penalty is adapted
  - penalized maximum likelihood finds the mode given the penalty parameter, and is almost the same as maximum a posteriori
  - when the amount of penalty is increased, marginal modes of weak effects go to zero first
  - when the amount of penalty is increased, also the relevant coefficients are shrunk towards zero
  - sometimes relaxed lasso is used, where after variable selection coefficients are re-estimated
- Bayesian lasso uses Laplace distribution as prior
  - Laplace prior is equivalent to L1 penalty
  - but the Bayesian inference includes distribution for parameters and that distribution doesn't shrink to a point at zero, even if the mode would be at zero
  - empirically better results obtained with more sparse priors
  - it's best to separate selection of sensible prior, good posterior inference, and the decision analysis of which variables are important

# Sparse priors




from Carvalho, Polson, Scott (2009).

# Regularized horseshoe



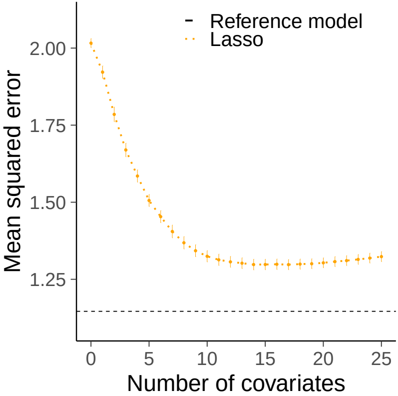

for more see

- Piironen and Vehtari (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. In Electronic Journal of Statistics, 11(2):5018-5051. Online
- https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html

# Projpred selection vs. Lasso

See projpred in lecture 9.3

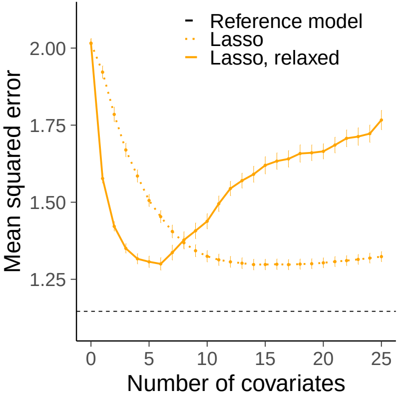
Same simulated regression data as in lecture 9,3,
$n = 50$, $p = 500$, $p_{\mathrm{rel}} = 150$, $\rho = 0.5$



# Projpred selection vs. Lasso

See projpred in lecture 9.3

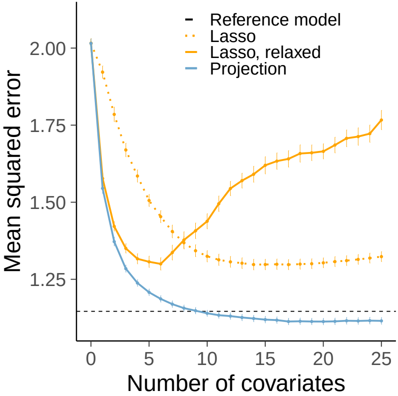
Same simulated regression data as in lecture 9,3,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



# Projpred selection vs. Lasso

See projpred in lecture 9.3

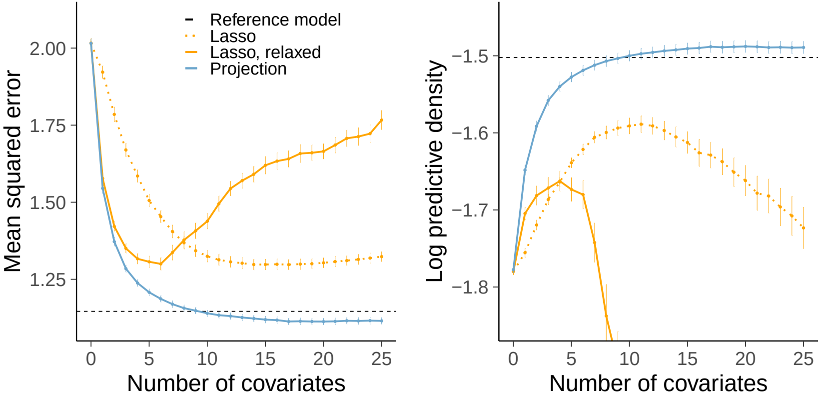
Same simulated regression data as in lecture 9,3,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



# Projpred selection vs. Lasso

See projpred in lecture 9.3

Same simulated regression data as in lecture 9,3,
$n = 50$, $p = 500$, $p_{\mathrm{rel}} = 150$, $\rho = 0.5$



## Chapter 15: Hierarchical linear models

- Since you know hierarchical models, theory is easy
- With probabilistic programming computation is also easy
  - BDA3 discusses some other computational issues
  - section on transformations for HMC is relevant
    (see also Stan user guide 21.7 Reparameterization)

## Chapter 15: Hierarchical linear models

- Since you know hierarchical models, theory is easy
- With probabilistic programming computation is also easy
  - BDA3 discusses some other computational issues
  - section on transformations for HMC is relevant
    (see also Stan user guide 21.7 Reparameterization)
- Fixed, random, and mixed effects models
  - we don't recommend using these terms, but they are so popular that it's useful to know them

# Chapter 15: Hierarchical linear models

- Since you know hierarchical models, theory is easy
- With probabilistic programming computation is also easy
    - BDA3 discusses some other computational issues
    - section on transformations for HMC is relevant (see also Stan user guide 21.7 Reparameterization)
- Fixed, random, and mixed effects models
    - we don't recommend using these terms, but they are so popular that it's useful to know them

| | |
|---|---|
| `y ~ 1 + x` | fixed / population effect; pooled model |
| `y ~ 1 + (0 + x | g)` | random / group effects |
| `y ~ 1 + x + (1 + x | g)` | mixed effects; hierarchical model |

- ANOVA in section 15.6 (see also `stan_aov`)

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
  1. The linear predictor $\eta = X\beta$

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
  1. The linear predictor $\eta = X\beta$
  2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
  1. The linear predictor $\eta = X\beta$
  2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
  3. Outcome distribution model with location parameter $\mu$

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
    1. The linear predictor $\eta = X\beta$
    2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
    3. Outcome distribution model with location parameter $\mu$
        - the distribution can also depend on dispersion parameter $\phi$

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
  1. The linear predictor $\eta = X\beta$
  2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
  3. Outcome distribution model with location parameter $\mu$
     - the distribution can also depend on dispersion parameter $\phi$
     - originally just exponential family distributions (e.g. Poisson, binomial, negative-binomial), which all have natural location-dispersion parameterization

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
  1. The linear predictor $\eta = X\beta$
  2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
  3. Outcome distribution model with location parameter $\mu$
     - the distribution can also depend on dispersion parameter $\phi$
     - originally just exponential family distributions (e.g. Poisson, binomial, negative-binomial), which all have natural location-dispersion parameterization
     - after MCMC made computation easy, GLM can refer to models where outcome distribution is not part of exponential family and dispersion parameter may have its own latent linear predictor

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
  1. The linear predictor $\eta = X\beta$
  2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
  3. Outcome distribution model with location parameter $\mu$
     - the distribution can also depend on dispersion parameter $\phi$
     - originally just exponential family distributions (e.g. Poisson, binomial, negative-binomial), which all have natural location-dispersion parameterization
     - after MCMC made computation easy, GLM can refer to models where outcome distribution is not part of exponential family and dispersion parameter may have its own latent linear predictor
  - Hierarchical GLM natural extension

# Chapter 16: Generalized linear models

- Bioassay model is an example of GLM
- Components:
    1. The linear predictor $\eta = X\beta$
    2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
    3. Outcome distribution model with location parameter $\mu$
        - the distribution can also depend on dispersion parameter $\phi$
        - originally just exponential family distributions (e.g. Poisson, binomial, negative-binomial), which all have natural location-dispersion parameterization
        - after MCMC made computation easy, GLM can refer to models where outcome distribution is not part of exponential family and dispersion parameter may have its own latent linear predictor
    - Hierarchical GLM natural extension
    - 16.3 Weakly informative priors section is excellent although the recommendation on using Cauchy has changed (see https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations)

# Chapter 17: Models for robust inference

- For example

| | | |
|---|---|---|
| normal | $\rightarrow$ | $t$-distribution |
| Poisson | $\rightarrow$ | negative-binomial |
| binomial | $\rightarrow$ | beta-binomial |
| probit | $\rightarrow$ | logistic / robit |

# Chapter 17: Models for robust inference

- For example

| | | |
|---|---|---|
| normal | $\rightarrow$ | $t$-distribution |
| Poisson | $\rightarrow$ | negative-binomial |
| binomial | $\rightarrow$ | beta-binomial |
| probit | $\rightarrow$ | logistic / robit |

- Computation with MCMC easy
  - posterior can be multimodal

# Chapter 17: Models for robust inference

- For example

| | | |
|---|---|---|
| normal | $\rightarrow$ | $t$-distribution |
| Poisson | $\rightarrow$ | negative-binomial |
| binomial | $\rightarrow$ | beta-binomial |
| probit | $\rightarrow$ | logistic / robit |

- Computation with MCMC easy
  - posterior can be multimodal
  - rstanarm doesn't have $t$-distribution for outcome, but brms has

## Chapter 18: Models for missing data

- Extends the data collection modelling from Chapter 8
- Useful terms

# Chapter 18: Models for missing data

- Extends the data collection modelling from Chapter 8
- Useful terms
  - Missing completely at random (MCAR)
    missingness does not depend on missing values or other observed values (including covariates)

## Chapter 18: Models for missing data

- Extends the data collection modelling from Chapter 8
- Useful terms
  - Missing completely at random (MCAR)
    missingness does not depend on missing values or other observed values (including covariates)
  - Missing at random (MAR)
    missingness does not depend on missing values but may depend on other observed values (including covariates)

# Chapter 18: Models for missing data

- Extends the data collection modelling from Chapter 8
- Useful terms
  - Missing completely at random (MCAR)
    missingness does not depend on missing values or other observed values (including covariates)
  - Missing at random (MAR)
    missingness does not depend on missing values but may depend on other observed values (including covariates)
  - Missing not at random (MNAR)
    missingness depends on missing values

# Chapter 18: Models for missing data

- Extends the data collection modelling from Chapter 8
- Useful terms
  - Missing completely at random (MCAR)
    missingness does not depend on missing values or other observed values (including covariates)
  - Missing at random (MAR)
    missingness does not depend on missing values but may depend on other observed values (including covariates)
  - Missing not at random (MNAR)
    missingness depends on missing values
- Multiple imputation
  1. make a model predicting missing data
  2. sample repeatedly from the missing data model to generate multiple imputed data sets
  3. make usual inference for each imputed data set
  4. combine results

# Chapter 21: Gaussian process models

- Gaussian process is
  - infinite dimensional extension of normal distribution
  - useful prior for non-linear functions
  - for any finite number of variables, the marginal is multivariate normal $f_1, \ldots, f_n \sim \mathrm{N}\left(\mu(x_1, \ldots, x_n), K(x_1, \ldots, x_n)\right)$

# Chapter 21: Gaussian process models

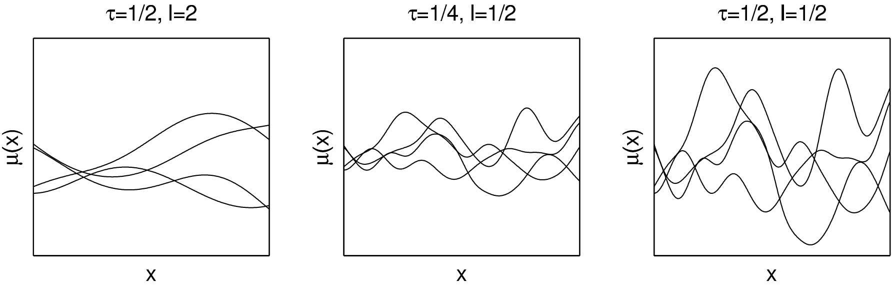
- Gaussian process is
  - infinite dimensional extension of normal distribution
  - useful prior for non-linear functions
  - for any finite number of variables, the marginal is multivariate normal $f_1, \ldots, f_n \sim \mathrm{N}\left(\mu(x_1, \ldots, x_n), K(x_1, \ldots, x_n)\right)$
- Often a priori $\mu = 0$

# Chapter 21: Gaussian process models

- Gaussian process is
  - infinite dimensional extension of normal distribution
  - useful prior for non-linear functions
  - for any finite number of variables, the marginal is multivariate normal $f_1, \ldots, f_n \sim \mathrm{N}\left(\mu(x_1, \ldots, x_n), K(x_1, \ldots, x_n)\right)$
- Often a priori $\mu = 0$
- Prior for smooth non-linear functions, e.g. with $k(x, x') = \tau^2 \exp\left(-\frac{|x-x'|^2}{2l^2}\right)$

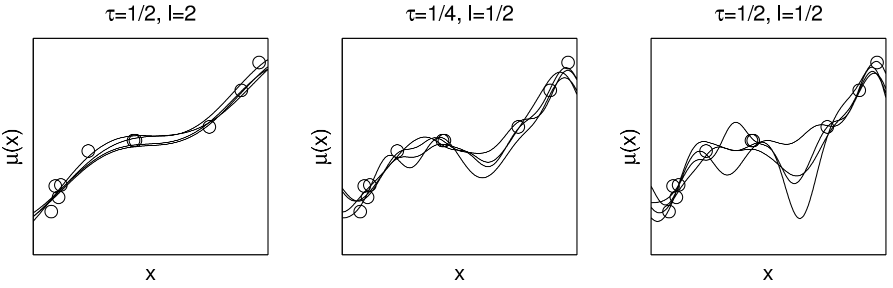
# Chapter 21: Gaussian process models

- Gaussian process is
  - infinite dimensional extension of normal distribution
  - useful prior for non-linear functions
  - for any finite number of variables, the marginal is multivariate normal $f_1, \ldots, f_n \sim \mathrm{N}\left(\mu(x_1, \ldots, x_n), K(x_1, \ldots, x_n)\right)$
- Often a priori $\mu = 0$
- Prior for smooth non-linear functions, e.g. with $k(x, x') = \tau^2 \exp\left(-\frac{|x-x'|^2}{2l^2}\right)$



# Chapter 21: Gaussian process models

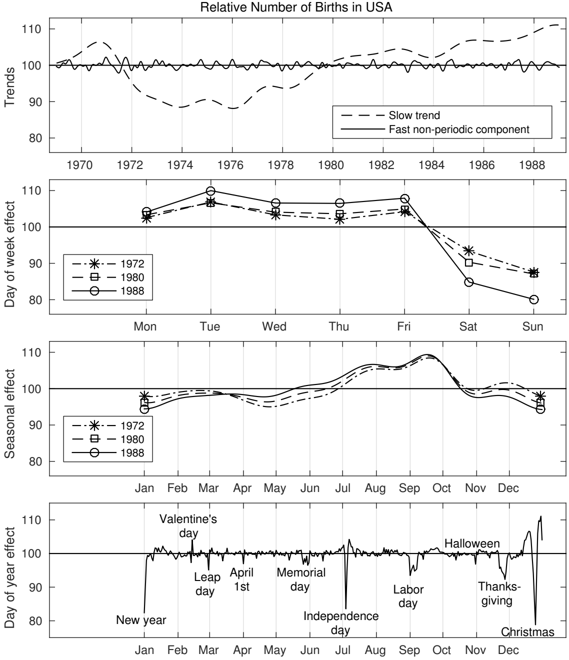
- Gaussian process is
  - infinite dimensional extension of normal distribution
  - useful prior for non-linear functions
  - for any finite number of variables, the marginal is multivariate normal $f_1, \ldots, f_n \sim \mathrm{N}\left(\mu(x_1, \ldots, x_n), K(x_1, \ldots, x_n)\right)$
- Often a priori $\mu = 0$
- Prior for smooth non-linear functions, e.g. with $k(x, x') = \tau^2 \exp\left(-\frac{|x-x'|^2}{2l^2}\right)$



## Chapter 21: Gaussian process models

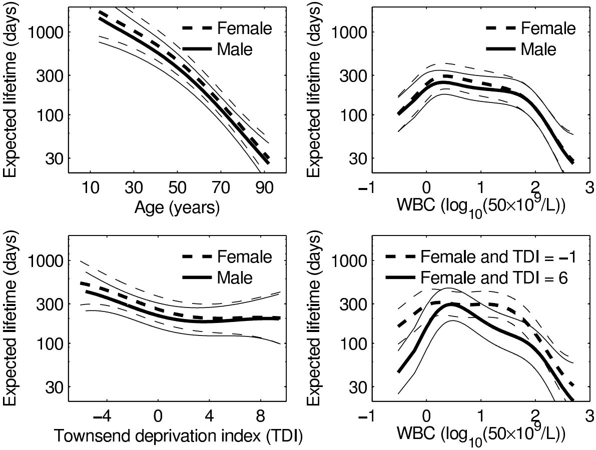
- Conditional on covariance function parameter the posterior is just multivariate normal
  - need to make inference for covariance function parameters given the marginal likelihood
  - the exact computation of the marginal likelihood scales $O(N^3)$

- Easy to make additive models
$y_t(t) = f_1(t) + f_2(t) + f_3(t) + f_4(t) + f_5(t) + \epsilon_t$



# Chapter 21: Gaussian process models

- For non-Gaussian outcome models similar extension as GLMs
- Survival model example:



## GPs in Stan

- GP specific software (e.g. GPy, GPflow, GPyTorch) scale computationally better for GPs than Stan
- Stan has some built-in covariance functions (and soon GPU support)
- In case of non-Gaussian outcome models, sampling of latent variables can be slow (Laplace integration over the latents coming)

## GPs in Stan

- GP specific software (e.g. GPy, GPflow, GPyTorch) scale computationally better for GPs than Stan
- Stan has some built-in covariance functions (and soon GPU support)
- In case of non-Gaussian outcome models, sampling of latent variables can be slow (Laplace integration over the latents coming)
- Instead of covariance matrix based approach, for low dimensional cases faster to use basis function representation
    - e.g. `stan_glm(y ~ s(x, bs="gp"))`