Ludwig-Maximilians-Universität München
Institut für Statistik

**Master-Thesis**

# Predictive Assessment of Bayesian Hierarchical Models

Daniel Sabanés Bové

September 2009

Supervised by
Prof. Ludwig Fahrmeir/Ludwig-Maximilians-Universität München
and Prof. Leonhard Held/Universität Zürich

# Abstract

Bayesian hierarchical models are increasingly used in many applications. In parallel, the desire to check the predictive capabilities of these models grows. However, classic Bayesian tools for model selection, as the marginal likelihood of the models, are often unavailable analytically, and the models have to be estimated with MCMC methodology. This also renders leave-one-out cross-validation of the models infeasible for realistically sized data sets. In this thesis we therefore propose approximate cross-validation sampling schemes based on work by Marshall and Spiegelhalter (2003), for two model classes: conjugate change point models are applied to time series, while normal linear mixed models are used to analyze longitudinal data. The quality of the models' predictions for the left-out data is assessed with calibration checks and proper scoring rules. In several case studies we show that the approximate cross-validation results are typically close to the exact cross-validation results, and are much better suited for predictive model assessment than analogous posterior-predictive results, which can only be used for goodness-of-fit checks.

# Preface

This master thesis was written between 1st April 2009 and 20th September 2009 at the Department of Statistics of the Ludwig-Maximilians-Universität München (Munich).

The board of examiners kindly allowed me to be supervised by Prof. Dr. Leonhard Held (University of Zurich), whom I would like to express my gratitude for his invaluable advice and constant support. Equally I would like to thank Prof. Dr. Ludwig Fahrmeir, who supervised the second part of the thesis. Part of this work was done while I was a student assistant at the Department of Statistics. I am also indebted to Andreas Bayerstadler, who provided me with a perfectly prepared version of the CD4 data set and thorough documentation of it. Felix Heinzl sent me the BMI data set. The seamless usage of the `BayesX` program would not have been possible without the help of Thomas Kneib. Last but not least I owe a big thank you to my parents.

This thesis has been compiled with LaTeX (Lamport 1999). All figures and most tables have been produced by means of the free statistical computing software `R` (R Development Core Team 2009), which has also been utilized for most computations. Parts of the computations were done with `C++` programs compiled with the GNU `C++` compiler (gcc 4.3.2, `http://gcc.gnu.org/`), and of course `BayesX` was used for the second part. The newly created software is assembled in a documented `R`-package, which is attached to the electronic version of the thesis as supplementary material. It can be installed and used in any operating system with `R`, `C++` compiler and `BayesX` installations. Since all results presented in the thesis were integrated using the literal programming platform `Sweave` (Leisch 2002), the reproducibility of the results should be guaranteed.

# Contents

# 1 Introduction

One of the major tasks of statistics is to issue forecasts for the future, based on evidence from the past (Dawid 1984). The evidence usually has the form of a data set which contains the target variable one wants to predict for the future (the response), and multiple variables known or suspected to influence the target in some way (the covariates). In model-based statistics, a stochastic model is fit to the known data set, which can then be used to predict unknown responses from the corresponding known covariates. If prediction is a major task in the application, the model's predictive capabilities must be assessed, in order to compare it with other models or to know how to improve it. Although this approach to statistical inference is not indisputable, the majority of the statistical discipline works with this scheme (Breiman 2001). The general problem is to find a model which fits the past data well enough to capture those relationships between covariates and response that are important for the prediction of future data, but does not over-interpret noise in the data set which could lead to prediction artifacts. Models which do not capture the important relationships suffer from "underfitting", while models over-interpreting noise suffer from "overfitting" of the data set. In particular, a model assessment which is only based on the goodness-of-fit of the model to the known data set will tend to favour overfitting models, while a too simple stochastic model could lead to underfitting.

A general tool for predictive assessment of statistical models is cross-validation. The most primitive form "consists in the controlled or uncontrolled division of the data sample into two subsamples, the choice of a statistical predictor, including any necessary estimation, on one subsample and then the assessment of its performance by measuring its predictions against the other subsample" (Stone 1974, p. 111). So we hide an actually known part of the past data from the model, to be able to compare its predictions with this pseudo-future data. A popular type of cross-validation which "squeezes the data almost dry" is leave-one-out cross-validation: "set aside one individual case, optimize for what is left, then test on the set-aside case", and repeat that for every case (Mosteller and Tukey 1968). Comparing different models assessed on the same data set, we can then choose the model which has the best cross-validation performance, with regard to an appropriate measure, and are thus protected from favouring overfitting models. Yang (2007) shows that under regularity assumptions, cross-validation is consistent for increasing sample size in the sense of selecting the better model with probability approaching 1.

While the generic concept of cross-validation is applicable to all estimation concepts, usually model assessment in Bayesian inference is done differently. The classic approach starts with expressing the prior model preferences as a prior distribution on the models, that is, without having looked in the data set, how probable is it that each model is the true model? Via Bayes' theorem, these probabilities are then updated to the posterior model probabilities by the information contained in the data set. Afterwards, the whole model assessment can be based on these probabilities. For example, one can choose the model with the highest posterior probability, or average the quantities of interest over models by weighting them with the posterior probabilities. Clyde and George (2004) give an overview of Bayesian treatment of model uncertainty. When the model prior is constant on the (finite) model space, the posterior model probabilities are proportional to the marginal likelihood values of the models. Even in this case where the model prior is indifferent to the complexity of the models, this approach is guarded against overfitting by the Bayesian "Ockham's Razor" (Jefferys and Berger 1992). The reason is that the marginal likelihood of a model, which is the value of the marginal density under this model at the observed data, rewards simple models for their sharp prediction if the observed data lies in their support. By contrast, more complex models spread their probability mass to larger regions, and thus have lower density values.

In recent years, proper scoring rules as another general tool for predictive assessment have become popular (Gneiting and Raftery 2007). Scoring rules assign a forecasting distribution a (penalty) score, based on a comparison with the materialized observation. The rule is (strictly) proper if the resulting expected score, with respect to the true data generating distribution, is (uniquely) optimized when the forecasting distribution is identical to the data generating distribution. This regularity requirement is necessary to force the scoring rule to prefer honest forecasts, by addressing both the sharpness and the calibration of the forecasts. It is also possible to separately assess the calibration, which can be summarized as the consistency between the forecast quantiles and the observed data quantiles (Gneiting, Balabdaoui, and Raftery 2007). Proper scoring rules are usually utilized as distance measures between predictive distributions and observations in cross-validation setups, where the model score is then defined as the average of the single scores for the test samples. In time series modelling, the one-step-ahead assessment, which iteratively enlarges the training part of the data with the next observation in time, is an alternative. We will see that in this case, the one-step-ahead validated model score obtained from the logarithmic scoring rule (which is the log of the predictive density evaluated at the materialized observation) and the marginal likelihood are equivalent.

The logarithmic scoring rule is also linked to Akaike's Information Criterion (AIC), which is often used to compare models estimated by maximum likelihood (Akaike 1974):

Stone (1977) shows that the leave-one-out cross-validated log-score of a model and AIC are asymptotically equivalent, with regard to an increasing size of the data set. This is comprehensible, as the AIC is defined as the maximized log-likelihood (that is, the log data density evaluated at the parameter estimate) penalized with the dimension of the parameter in the model – so the AIC definition contains the log of the full data density. A similar form has the Deviance Information Criterion (DIC), which was proposed by Spiegelhalter, Best, Carlin, and van der Linde (2002) as a Bayesian measure for both model complexity and fit: It penalizes the posterior expected deviance with an estimate of the effective number of parameters in the Bayesian model (see appendix A.2 for the details). The DIC can be estimated with posterior parameter samples from obtained from Markov chain Monte Carlo (MCMC) methods, so that it can also be estimated if the marginal likelihood of a Bayesian model is not analytically available. Another criterion, the Bayesian Information Criterion (BIC), is asymptotically equivalent to the marginal likelihood (Schwarz 1978). The BIC is similar to the AIC, but weights the parameter dimension with the log of the number of observations in the data set, leading to a stronger penalization of the maximized log-likelihood; see Kuha (2004) for a good comparison of AIC and BIC.

The DIC is especially popular for the assessment of Bayesian hierarchical models, i. e. models with multiple layers of parameters, which are estimated within the Bayesian inference framework. In this thesis we want to do cross-validation of two special types of Bayesian hierarchical models, where we measure the quality of the predictions for the left out observations by proper scoring rules or calibration checks. Because the models are estimated with computationally intensive Monte Carlo algorithms, the exact cross-validation will only be feasible for small sample sizes. Thus, we follow Mosteller and Tukey (1968), who further write:

> "If we have to go through the full optimization calculation every time, the extra computation may be hard to face. Occasionally we can easily calculate [. . .] to an adequate approximation what the effect of dropping a specific and very small part of the data will be on the optimized result. [. . .] That is, we make one optimization for all the data, followed by one repetition per case of a much simpler calculation, a calculation of the effect of dropping each individual, followed by one test of that individual. When practical, this approach is attractive."

Except that we will draw samples from the parameter posterior instead of optimizing the parameter of the model, this is exactly what we will do in our approximate cross-validations, where the approximation is based on work by Marshall and Spiegelhalter (2003). We will investigate in case studies how good these approximations are, how much computing time they save and what the effect is on the model choice.

The outline of the thesis is as follows. In chapter 2, we describe the tools (proper scoring rules and calibration checks) for evaluation of the predictive distributions with respect to the materialized observations, which are used in the following two chapters. Chapter 3 examines conjugate change point models, which are useful for time series modelling. For three distribution families, exact and approximate predictive assessment are compared, before the approximate approach is applied to a genetic data set. Chapter 4 examines random effects models for longitudinal data. For two real data sets, the exact and approximate approach are first compared on a subset of feasible size, before the cross-validation is approximated on the whole data set. A simulation study with known true model will yield interesting results. The thesis findings are summarized and discussed in chapter 5.

# 2 Evaluating predictive distributions

Section 2.1 introduces the setting and nomenclature for this chapter. Tools for assessing the probabilistic calibration of predictive distributions are described in section 2.2. The other type of tools for evaluating forecasters in this thesis are proper scoring rules, which are presented in section 2.3. An outlook on the application of custom summary statistics for tests of specific aspects of predictive distributions is given in section 2.4.

## 2.1 Introduction

This chapter describes techniques for evaluating predictive distributions with respect to the materialized observation which has been predicted. The predictive distributions can belong to probabilistic forecasts of a future observation, but might also be posterior-predictive distributions for a known observation – the origin of the predictive distribution is not of interest in this chapter. This is in accordance with the Prequential Principle of Dawid (1984, p. 281).

Starting with the univariate case, we assume that the predictive distribution has cumulative distribution function (cdf) $F$ and denote the prediction random variable by $Y$. That is, $Y \sim F$. Since our diagnostic tools will be based on Monte Carlo estimates, assume that $m$ independent identically distributed (iid) samples from $F$ are available:

$$y_{[j]} \overset{iid}{\sim} F, \quad j = 1, 2, \dots, m.$$

The empirical cdf of this sample of size $m$ is $\hat{F}_m(y) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{[y_{[j]}, +\infty)}(y)$. The materialized observation is $x$. It is a realization of the random variable $X$ with cdf $G$, thus $X \sim G$.

In the multivariate case, we want to predict a vector-valued observation $\boldsymbol{x} \in \mathbb{R}^k$. It is a realization of the random vector $\boldsymbol{X}$. The prediction random vector is $\boldsymbol{Y} : \Omega \to \mathbb{R}^k$, and we again assume that $m$ iid realizations $\boldsymbol{y}_{[1]}, \dots, \boldsymbol{y}_{[m]}$ of $\boldsymbol{Y}$ are available.

The methodology is based on comparing a single predictive distribution with the corresponding materialized observation. In practice however and in our applications in chapters 3 and 4, there will be multiple observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and corresponding predictive cdfs $F_1, \dots, F_n$, which shall be evaluated together. This requirement will also be considered in the following sections.

## 2.2 Assessing probabilistic calibration

Gneiting, Balabdaoui, and Raftery (2007, p. 247) define probabilistic calibration by means of the countable sequences $(F_i)_{i \in \mathbb{N}}$ of predictive cdfs and $(G_i)_{i \in \mathbb{N}}$ of corresponding (in practice unknown) true observation cdfs: if for all probabilities $p \in (0, 1)$ the average $\frac{1}{n} \sum_{i=1}^{n} G_i\big(F_i^{-1}(p)\big)$ converges almost surely to $p$ when the number of instances $n \to \infty$, then $(F_i)$ is probabilistically calibrated relative to $(G_i)$. (The stochastic notion of convergence is needed because the cdfs might depend on stochastic parameters.) We will drop the word “probabilistic” in the term from now on, because we only consider this mode of calibration.

Note that if $F_i \equiv G_i$ for all instances $i$, then the predictions are trivially calibrated. We call a predictive distribution which is identical to the unknown true data generating distribution the ideal forecaster. However, the upper definition is only sensible if the cdfs are continuous and invertible. In general, calibration is best described as “the statistical consistency between the distributional forecasts and the observations, and is a joint property of the forecasts and the events or values that materialize” (Gneiting and Raftery 2007, p. 359). With the Probability Integral Transform and the Box Ordinate Transform we present tools which can be used for the comparison of the predictions $F_i$ and (possibly vector-valued) observations $\boldsymbol{x}_i$ (instead of the unknown $G_i$), to assess the calibration of the predictions $F_i$.

### 2.2.1 Probability Integral Transform

The Probability Integral Transform (PIT) was introduced by Dawid (1984, p. 281). It is defined as

$$PIT(F, x) := F(x), \tag{2.2.1}$$

with the notation emphasizing that the PIT value depends on both the predictive cdf $F$ and the value $x$ that materializes. The PIT is only useful for univariate observations $x$.

If the predictive distribution $F$ matches the data generating distribution of a continuous random variable $X$ exactly, then it is well-known that $PIT(F, X) = F(X) \sim \mathrm{U}(0, 1)$ (Gneiting, Balabdaoui, and Raftery 2007, p. 244). Given an independent sample $x_1, \dots, x_n$ with corresponding predictive distributions $F_1, \dots, F_n$, the empirical distribution $H$ of the PIT values $F_1(x_1), \dots, F_n(x_n)$ can be compared against the standard uniform distribution. For that purpose, usually a PIT histogram is plotted. If the $F_i$ cannot be evaluated analytically, they can be estimated by empirical cdfs $\hat{F}_{i,m}$, using samples $y_{i,[1]}, \dots, y_{i,[m]}$ from the distributions $F_i$. It can be shown (Gneiting, Balabdaoui, and Raftery 2007, p. 252) that the (almost sure) convergence in $n \to \infty$ of the PIT histogram to the density histogram of the uniform distribution is equivalent to the original definition given above.

Characteristic deviations of the PIT histograms from uniformity can point out uncalibrated predictive densities, which is illustrated by histograms obtained from normal data generating and forecasting distributions in Figure 2.1 on page 10. Here no Monte Carlo estimation of the tail probability is necessary, since

$$\mathbb{P}(Y \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

when $Y \sim \mathrm{N}(\mu, \sigma^2)$ is the prediction random variable. The u-shaped form of the PIT histogram in panel (a) is typical for underdispersed predictive distributions. This form is understandable from the PIT definition (2.2.1): the data generating density has heavier tails than the forecaster, and when an extreme observation is generated, the PIT value is either very low or very high. The hump-shaped form in panel (c) is typical for overdispersed predictive distributions, and can be explained similarly. A uniform PIT histogram as in panel (b) is expected for the ideal forecaster.

For discrete random variables $X_i$, the PIT values are no longer distributed uniformly, so an ordinary PIT histogram will look strange even if the predictive distribution is perfectly calibrated. For count data and ordered categorical data, Czado, Gneiting, and Held (2009) have proposed a modified PIT histogram. The idea can be understood quickly in terms of the histogram-generating distribution function $H$. For continuous variables $X_i$ with realizations $x_i$, $H(y)$ is the average of discontinuous indicator functions $\mathbb{I}_{[F_i(x_i),+\infty)}(y)$ over the observations $i = 1, \dots, n$. So the distribution $H$ is a mixture of the $n$ point-masses $\delta_{F_i(x_i)}$, $i = 1, \dots, n$. For count variables with support $\mathbb{N}_0$, Czado et al. (2009) define $H$ as the mixture of the $n$ continuous uniform distributions $\mathrm{U}\big(F_i(x_i - 1), F_i(x_i)\big)$ over the observations $i = 1, \dots, n$. The resulting distribution $H$ is thus always continuous, and is expected to be the standard uniform distribution for perfectly calibrated $F_i$, in the sense that $\mathbb{E}_X\, H(y) = y$ for any $y \in [0, 1]$. This can again be checked by plotting a density histogram of $H$. Recalling that the PIT values are special p-values, we can use the mid-p-values

$$\mathbb{P}(Y_i < x_i) + \frac{1}{2}\,\mathbb{P}(Y_i = x_i)$$

which have been used e. g. by Marshall and Spiegelhalter (2003, p. 1651). For count variables, these are the same as the midpoints of the uniform distributions supports in the modified PIT histogram, namely

$$\frac{1}{2}\big(F_i(x_i - 1) + F_i(x_i)\big).$$

### 2.2.2 Box Ordinate Transform

The Box Ordinate Transform (BOT) was introduced by Box (1980, p. 386) for the special case of Bayesian estimation of the mean of a normal distribution. It is defined as the tail probability

$$BOT(f, \boldsymbol{x}) := \mathbb{P}\big(f(\boldsymbol{Y}) \leq f(\boldsymbol{x})\big) = \int \mathbb{I}_{\{\boldsymbol{z}: f(\boldsymbol{z}) \leq f(\boldsymbol{x})\}}(\boldsymbol{y}) f(\boldsymbol{y})\, d\boldsymbol{y}, \tag{2.2.2}$$

where $f$ is the continuous Lebesgue density of $\boldsymbol{Y}$. The BOT has strong connections to significance (and especially likelihood ratio type) tests: Assuming that it is really the density $f$ which produces the observation $\boldsymbol{x}$, what is the probability of observing an even smaller density ordinate than the observed $f(\boldsymbol{x})$? The BOT can also be used for univariate observations, but it is the only adequate calibration checking tool (from those introduced in this thesis) for multivariate observations (leaving aside the multivariate rank and minimum spanning tree rank histograms from Gneiting, Stanberry, Grimit, Held, and Johnson (2008, p. 215), for example).

The BOT was hence used as a model checking tool, where $f$ was the prior predictive density under the assumed model. For example, Sinharay and Stern (2003, p. 214) call it "the prior predictive method of Box" and stress that it could only be used if the parameters prior in the assumed model was proper, as otherwise the prior predictive density would not exist. Here, however, we are sure that our predictive distribution $F$ with density $f$ exists and we have available samples $\boldsymbol{y}_{[1]}, \dots, \boldsymbol{y}_{[m]}$ from $F$, so this critique need not concern us.

From another point of view we can easily see that $BOT(f, \boldsymbol{X})$ has a uniform distribution whenever $\boldsymbol{X}$ really has probability density $f$, as stated by Gneiting, Stanberry, Grimit, Held, and Johnson (2008, p. 220): Consider the scalar random variable $Z := f(\boldsymbol{Y})$ as the transformation of $\boldsymbol{Y} \sim f$ onto the positive real line, with cdf $F_Z$. Thus, obviously $BOT(f, \boldsymbol{x}) = F_Z(f(\boldsymbol{x}))$. If $\boldsymbol{X} \sim f$, then $f(\boldsymbol{X})$ is identically distributed to $Z$ and has cdf $F_Z$, and $BOT(f, \boldsymbol{X})$ is identically distributed to $F_Z(Z) = PIT(F_Z, Z)$. So in fact, the BOT is a PIT value on the predictive density scale! If $Z$ is a continuous random variable, then this raw PIT value is uniform (cf. page 6). In our applications, this condition will be satisfied because the predictive density $f$ will always be a Lebesgue density without plateaus. This ensures that given $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \overset{iid}{\sim} f$ the probability of $Z_1 = f(\boldsymbol{Y}_1)$ and $Z_2 = f(\boldsymbol{Y}_2)$ being identical is zero.

The last question is how we estimate the BOT value $BOT(f, \boldsymbol{x})$. After having available the ordinate values $z_{\boldsymbol{x}} = f(\boldsymbol{x})$ and $z_{[1]} = f(\boldsymbol{y}_{[1]}), \dots, z_{[m]} = f(\boldsymbol{y}_{[m]})$, we could proceed as for the univariate PIT estimation, i. e. estimate the BOT by the empirical distribution function value

$$\widehat{BOT}(f, \boldsymbol{x}) := \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{[z_{[j]}, +\infty)}(z_{\boldsymbol{x}}). \tag{2.2.3}$$

Yet, in our applications the predictive density function $f$ is unknown. Thus, the ordinate values must be estimated. The estimates are also needed for the logarithmic score and the procedure is described in the corresponding section 2.3.3. Finally, the empirical distribution of the BOT values for all prediction locations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ can be compared to the uniform distribution. This check for calibration of the respective predictive densities $f_1, \dots, f_n$ is usually done using histograms, analogously to the PIT histograms.

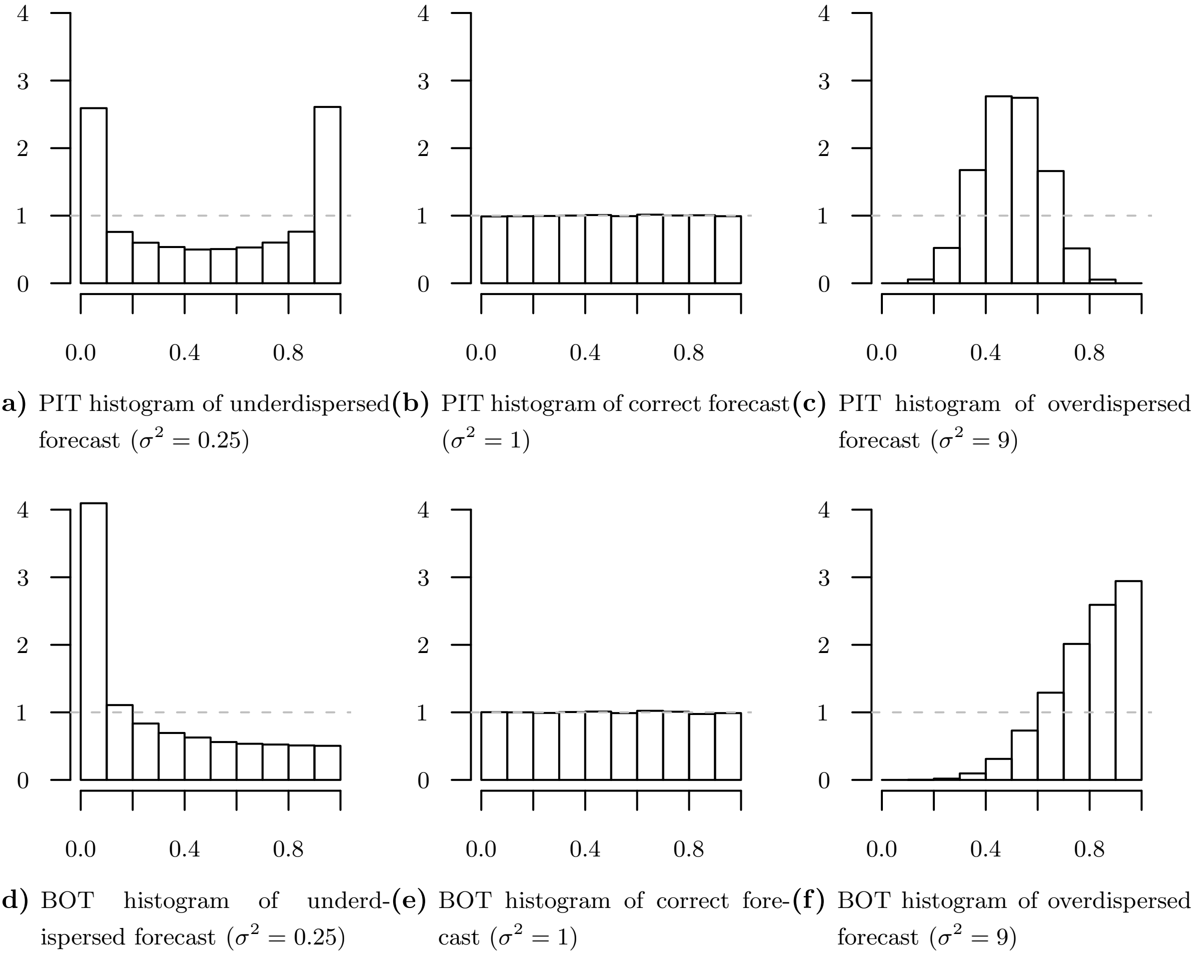
Characteristic deviations of the BOT histograms from uniformity can point out uncalibrated predictive densities, which is nicely illustrated in Figure 5 of Gneiting et al. (2008).[i] We show similar histograms obtained from normal data generating and forecasting distributions in Figure 2.1. Here no Monte Carlo estimation of the tail probability is necessary, since

$$\begin{aligned}
\mathbb{P}\big(f(Y) \leq f(x)\big) &= \mathbb{P}\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y-\mu)^2\right\} \leq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\right) \\
&= \mathbb{P}\left(\left(\frac{Y-\mu}{\sigma}\right)^2 \geq \left(\frac{x-\mu}{\sigma}\right)^2\right) \\
&= 1 - \chi^2\left(\left(\frac{x-\mu}{\sigma}\right)^2\right),
\end{aligned}$$

because $(Y-\mu)/\sigma \sim \mathrm{N}(0,1)$ when $f(y) = \mathrm{N}(y \,|\, \mu, \sigma^2)$ is the forecast density. The typical right-skewed BOT histogram for underdispersed forecasts is given in panel (d). The form can be understood from the BOT definition (2.2.2): since the data generating distribution puts large probability mass on areas where the forecast density has very low values, we often see small BOT values. On the other hand, BOT histograms for overdispersed forecasts are typically left-skewed as in panel (f). When the forecast is identical to the data generating distribution, we expect a uniform BOT histogram, here in panel (e).

However, it must be stressed that the uniform distribution of the BOT values is only a necessary, but not a sufficient condition for the calibration of a univariate forecast density. This is because the BOT is a PIT on the forecast density scale, and not on the original scale. A simple example which fulfills the regularity assumptions from above is described in the following. Let the observation random variable $X$ be beta-distributed $X \sim \mathrm{Be}(2,2)$, and define the forecast random variable $Y := \pm Z$ which switches the sign of the correct forecast $Z \sim \mathrm{Be}(2,2)$ with probability $1/2$. More formally, this is $Y := V \cdot Z$

[i] Note that their definition of the BOT on page 220 contains an error (T. Gneiting, personal communication), and our definition (2.2.2) is correct.



**(a)** PIT histogram of underdispersed forecast ($\sigma^2 = 0.25$)

**(b)** PIT histogram of correct forecast ($\sigma^2 = 1$)

**(c)** PIT histogram of overdispersed forecast ($\sigma^2 = 9$)

**(d)** BOT histogram of underdispersed forecast ($\sigma^2 = 0.25$)

**(e)** BOT histogram of correct forecast ($\sigma^2 = 1$)

**(f)** BOT histogram of overdispersed forecast ($\sigma^2 = 9$)

***Figure 2.1*** – *Simulation study for the PIT (upper row) and BOT (lower row) histograms. For each histogram, $n = 100\,000$ standard normal observations have been simulated. The density forecast is $\mathrm{N}(0, \sigma^2)$ with different variances $\sigma^2 = 0.25, 1, 9$ (columns).*

with $V \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$. If $g(x) = f_Z(x) = \mathrm{Be}(x \,|\, 2, 2)$ is the true data generating density, then we have the density $f_Y(y) = \frac{1}{2}g(|y|)$ for the wrong forecaster $Y$. Its BOT value for an observation $x \in (0, 1)$ is

$$\begin{aligned} BOT(f_Y, x) &= \mathbb{P}\big(f_Y(Y) \leq f_Y(x)\big) \\ &= \mathbb{P}\big(\tfrac{1}{2}g(|Y|) \leq \tfrac{1}{2}g(x)\big) \\ &= \mathbb{P}\big(g(Z) \leq g(x)\big) \\ &= BOT(g, x), \end{aligned}$$

because $|Y| = |V \cdot Z| = |V| \cdot Z = Z$. So the BOT value of the wrong forecaster $Y$ is identical to the BOT value of the correct forecaster $Z$, and the distribution of the BOT values is hence uniform also for the wrong forecaster $Y$. The corresponding PIT on the other hand is not uniform, because all PIT values $PIT(F_Y, x)$ must be greater than 1/2. Nevertheless, a non-uniform BOT histogram can be useful for the model critique, and should be used for random vectors because a multivariate PIT analogue would not be uniform for calibrated forecasts (Genest and Rivest 2001).

## 2.3 Proper scoring rules

In the last section, we have introduced the PIT and BOT histograms as tools for assessing the calibration of predictive distributions. However, these tools cannot distinguish every forecaster from the ideal forecaster. We have already given an example for the BOT, and Gneiting, Balabdaoui, and Raftery (2007, p. 244) give an example for the PIT, which we reiterate now. Let the true distribution be $G_i = \mathrm{N}(\mu_i, 1)$, the standard normal distribution shifted by $\mu_i \overset{iid}{\sim} \mathrm{N}(0, 1)$, $i \in \mathbb{N}$. The sequence of predictive distributions $F_i = \mathrm{N}(0, 2)$ which is constant over all times $i$ is then perfectly calibrated and cannot be distinguished by the PIT from the ideal forecaster $G_i$. This is easily seen from the marginal distribution of the observations $X_i$: Because the conditional distribution $X_i \,|\, \mu_i \sim \mathrm{N}(\mu_i, 1)$ is continuously mixed by $\mu_i \sim \mathrm{N}(0, 1)$, we have the marginal distribution $X_i \sim \mathrm{N}(0, 2)$ if we do not know the shifts $\mu_i$ (see appendix A.4 for the short proof). Thus $X_i$ is identically distributed as the prediction random variable $Y_i$, and $PIT(F_i, X_i) = \mathbb{P}(Y_i \leq X_i) \sim \mathrm{U}(0, 1)$. This example can be well summarized as the PIT-equivalence of the climatological forecaster $F_i$ and the ideal/meteorological forecaster $G_i$, which conditions on the current weather $\mu_i$ to predict the temperature $X_i$.

So we need other tools which help distinguishing calibrated forecasters from the ideal forecaster. Gneiting, Balabdaoui, and Raftery (2007, p. 245) propose the paradigm of “maximizing the sharpness of the predictive distributions subject to calibration”, where sharpness means the precision (literally the inverse variance if it exists) of the predictive distributions. In the example, this rule would prefer the ideal forecaster $G_i$ with precision 1 over the climatological forecaster $F_i$ with precision $\frac{1}{2}$. Yet, the direct assessment of sharpness is problematic in practice, because the PIT histograms will be different for all forecasters – then how should we combine this with some sharpness measure, e. g. the sharpness diagrams from Gneiting, Balabdaoui, and Raftery (2007, p. 261)? Moreover, for discrete distributions, these tools have not proved to be as useful as for continuous distributions (Czado, Gneiting, and Held 2009, p. 4).

This is where the proper scoring rules help us, as they can be used for an omnibus evaluation of both sharpness and calibration of predictive distributions. If the predictive distribution was chosen as $F$, and the observation $x$ materializes, the penalty score $S(F, x)$ is assigned by the scoring rule $S$. With the expected score under the data generating distribution $G$ being denoted as $S(F,G) = \int S(F,x)\,dG(x)$, a strictly proper scoring rule $S$ ensures that $S(G,G) \leq S(F,G)$ for all forecasters $F$ and $S(G,G) = S(F,G)$ if and only if $F = G$ (Gneiting and Raftery 2007, p. 359). We drop the adjective "strictly" because all used scoring rules in this thesis will be strictly proper. The propriety ensures that the ideal forecaster $G$ is preferred over all other forecasters, and that both sharpness and calibration of the forecaster are condensed into a single score (Winkler 1996).

In practice, we often want to combine scores $S(F_i, x_i)$, $i = 1, \dots, n$, into an overall score, which assesses the predictive performance of somehow comparable forecasters $F_1, \dots, F_n$ simultaneously. Then the mean score

$$\overline{S(F,x)} := \frac{1}{n} \sum_{i=1}^{n} S(F_i, x_i)$$

can be computed (Gneiting and Raftery 2007, p. 360). For example, the predictive distributions $F_i$ might come from the same parametric model. Then the comparison with another parametric model, producing forecasters $E_i$, say, is based on its mean score $\overline{S(E,x)}$. This procedure is theoretically well-founded, because the propriety of the single scores $S(F_i, x_i)$ ensures that the minimum of the mean score functional $\overline{S(\cdot,x)}$ is $\overline{S(G,x)}$ where $G$ denotes the sequence $G_1, \dots, G_n$ of true data generating distributions. That is, the mean score of single proper scores is again proper. When a formal statistical test for the comparison of $\overline{S(F,x)}$ and $\overline{S(E,x)}$ shall be utilized, a permutation test can be used. The details are given in appendix A.3.

We will use three specific scoring rules: The (continuous) ranked probability score is abbreviated as (C)RPS and can be used for univariate distributions only (section 2.3.1), while the energy score (ES) is the multivariate generalization (section 2.3.2). The logarithmic score (log-score) can be used for scalar and vector-valued observations, and needs predictive density evaluations instead of expectations under the predictive distributions (section 2.3.3).

### 2.3.1 Continuous ranked probability score

The general CRPS is defined as

$$CRPS(F,x) := \int_{-\infty}^{+\infty} \left\{F(y) - \mathbb{I}_{[x,+\infty)}(y)\right\}^2 dy, \tag{2.3.1}$$

which is the squared $L_2$-distance of the cdfs of the predictive distribution $F$ and the point-mass $\delta_x$ in the realized observation $x$, respectively. The CRPS is (strictly) proper if one considers predictive distributions $F$ with finite expectation (Gneiting and Raftery 2007, p. 367).

We want to use the $m$ samples $y_{[1]}, \dots, y_{[m]}$ from $F$ to estimate (2.3.1). The simplest idea is to replace the not analytically available cdf $F$ with the consistent estimate $\hat{F}_m$, and use $\widehat{CRPS}_m(F, x) := CRPS(\hat{F}_m, x)$. The evaluation of the latter integral is easy, because the integrand is a step function, with jumps at $x$ and at the quantiles of the sample.

For a continuous distribution $F$, the sample values are (almost surely) unique, so that the order statistic $y_{(1)}, y_{(2)}, \dots, y_{(m)}$ is a permutation of the original sample. Assume that the materialized observation is between $y_{(k-1)}$ and $y_{(k)}$ for some $k \in \{2, \dots, m\}$. The formula

$$\begin{aligned} CRPS(\hat{F}_m, x) = &\sum_{j=2}^{k-1} (y_{(j)} - y_{(j-1)}) \left(\frac{j-1}{m}\right)^2 + (x - y_{(k-1)}) \left(\frac{k-1}{m}\right)^2 \\ &+ (y_{(k)} - x) \left(\frac{m-(k-1)}{m}\right)^2 + \sum_{j=k+1}^{m} (y_{(j)} - y_{(j-1)}) \left(\frac{m-(j-1)}{m}\right)^2 \end{aligned}$$

is then derived straightforwardly. If $x < y_{(1)}$ or $x > y_{(m)}$, analogous formulae could be written down, and further illustration can be found in Hersbach (2000, p. 563).

For a count distribution $F$ with support $\mathbb{N}_0$, the cdfs can only jump at integer values. Then the ranked probability score

$$RPS(F, x) = \sum_{k \in \mathbb{N}_0} \left\{F(k) - \mathbb{I}_{[x,+\infty)}(k)\right\}^2$$

is derived from (2.3.1), cf. Czado, Gneiting, and Held (2009, section 3.2). If $\hat{F}_m(k)$ is the relative frequency of the samples less or equal to $k \in \mathbb{N}_0$, the estimator is

$$RPS(\hat{F}_m, x) = \sum_{k=\min\{y_{(1)}, x\}}^{\max\{y_{(m-1)}, x\}} \left\{\hat{F}_m(k) - \mathbb{I}_{[x,+\infty)}(k)\right\}^2 .$$

### 2.3.2 Energy score

The energy score (ES) can be applied to the prediction of multivariate quantities $\boldsymbol{x} \in \mathbb{R}^k$. It was proposed by Gneiting and Raftery (2007, p. 367) and is defined as

$$ES(F, \boldsymbol{x}) := \mathbb{E}\,\|\boldsymbol{Y} - \boldsymbol{x}\| - \frac{1}{2}\,\mathbb{E}\,\|\boldsymbol{Y} - \boldsymbol{Y}^*\|\,, \tag{2.3.2}$$

where $\boldsymbol{Y}, \boldsymbol{Y}^* \overset{iid}{\sim} F$ and $\|\boldsymbol{z}\|$ denotes the Euclidean norm $(\sum_{j=1}^{k} z_j^2)^{1/2}$ of $\boldsymbol{z} \in \mathbb{R}^k$. For dimension $k = 1$, it can be shown that

$$ES(F, x) = \mathbb{E}\,|Y - x| - \frac{1}{2}\,\mathbb{E}\,|Y - Y^*| = \int_{-\infty}^{+\infty} \left\{F(y) - \mathbb{I}_{[x,+\infty)}(y)\right\}^2 dy = CRPS(F, x),$$

meaning that the ES is a generalization of the CRPS for dimensions $k > 1$. The proof of the identity is detailed in appendix A.1. The ES is (strictly) proper if one considers predictive distributions $F$ with finite expectation (Gneiting and Raftery 2007, p. 367). (This assumption is also necessary for the identity of ES and CRPS.)

We want to use the $m$ samples $\boldsymbol{y}_{[1]}, \dots, \boldsymbol{y}_{[m]}$ from $F$ to estimate (2.3.2). An efficient Monte Carlo estimate proposed by Gneiting, Stanberry, Grimit, Held, and Johnson (2008, p. 223) is

$$\widehat{ES}(F, \boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^{m} \|\boldsymbol{y}_{[j]} - \boldsymbol{x}\| - \frac{1}{2(m-1)} \sum_{j=1}^{m-1} \|\boldsymbol{y}_{[j+1]} - \boldsymbol{y}_{[j]}\|,$$

where the computational cost is $O(m)$. If all pairwise Euclidean distances of the samples were utilized for the estimation of the expected between-forecasts distance, it would be $O(m^2)$. The precision of the estimator, however, would not be greatly increased, because the pairwise distances are not independent of each other.

### 2.3.3 Logarithmic score

Let $f$ be the (general) density of the predictive distribution $F$. The logarithmic score is then defined as

$$LogS(F, x) = -\log f(x), \tag{2.3.3}$$

where smaller score values are assigned to better predictive distributions. The logarithmic scoring rule is (strictly) proper both for discrete distributions (Gneiting, Balabdaoui, and Raftery 2007, p.352) and for continuous distributions when only forecasters with finite expectation are considered (Gneiting, Balabdaoui, and Raftery 2007, p.365).

For this score, samples from $F$ could only be used for nonparametric density estimation of $f$, which is often unstable. Yet, often and also in our applications the unknown density $f(x)$ is a continuous mixture of known densities $f(x \mid \theta)$,

$$f(x) = \int f(x \mid \theta) f(\theta)\, d\theta,$$

and we can produce samples $\theta_{[1]}, \dots, \theta_{[m]} \overset{iid}{\sim} f(\theta)$. Then the Monte Carlo estimate

$$\hat{f}(x) := \frac{1}{m} \sum_{j=1}^{m} f(x \mid \theta_{[j]}) \tag{2.3.4}$$

is preferable to a kernel density estimate which uses directly the samples $y_{[1]}, \dots, y_{[m]}$ which have been drawn from the conditional densities $f(y \mid \theta_{[1]}), \dots, f(y \mid \theta_{[m]})$. The formal justification for the superiority of the Monte Carlo estimate is based on the Rao-Blackwell theorem, see Gelfand and Smith (1990, p. 402). Yet, this estimate can be considered a special kernel density estimate where the kernels are the conditional densities, instead of the usual Gauss or Epanechnikov kernels (Davison 2003, p. 310).

In special cases, the logarithmic score can be computed analytically, as we will see for the one-step-ahead and leave-one-out scores in the conjugate change point model in section 3.3. The above Monte Carlo estimation is more often applicable, and will turn out to be very accurate.

## 2.4 Custom summary statistics

The evaluation of predictive distributions in this thesis will be based on the tools introduced in sections 2.2 and 2.3. However, there are many alternative proposals in the literature, which are often tailored to posterior-predictive model checking. See the references on page 190 in Gelman, Carlin, Stern, and Rubin (2003) for a good overview of the literature. We just try to sketch some of the popular ideas here, if possible for general predictive distributions.

One idea is to compute a scalar test statistic $T(\boldsymbol{x})$ of the observed data vector $\boldsymbol{x} \in \mathbb{R}^k$. The test statistic is chosen "to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied" (Gelman, Carlin, Stern, and Rubin 2003, p. 172). For example, in a longitudinal data setting, this could be the maximum, minimum or range of the data points $x_1, \dots, x_k$. The value $T(\boldsymbol{x})$ can then be compared with the distribution of the predicted test statistic, $T(\boldsymbol{Y})$. Usually some form of p-value is computed, which corresponds to the PIT value from section 2.2.1. Note that the BOT (2.2.2) fits in this framework with the test statistic $T$ being the predictive density $f$, such that the test statistic depends on the assumed model. However, we could also use the CRPS to judge the compatibility of $T(\boldsymbol{Y})$ and $T(\boldsymbol{x})$. Since the CRPS estimation in section 2.3.1 is based on samples, we just transform the original samples $\boldsymbol{y}_{[1]}, \dots, \boldsymbol{y}_{[m]}$ with $T$ to get the required scalar samples of the predicted test statistic.

A related concept are discrepancy measures $T(\boldsymbol{x}, \boldsymbol{\theta})$ which also depend on the assumed model through the parameter $\boldsymbol{\theta}$. Then tail probabilities of the form

$$\mathbb{P}\big(T(\boldsymbol{Y}, \boldsymbol{\Theta}) \geq T(\boldsymbol{x}, \boldsymbol{\Theta})\big)$$

are computed. For example, for a posterior-predictive check $\boldsymbol{\theta}_{[b]}$ is drawn from the posterior distribution, and $\boldsymbol{y}_{[b]}$ is drawn from the implied likelihood $f(\boldsymbol{y} \,|\, \boldsymbol{\theta}_{[b]})$, for $b = 1, \dots, B$. Afterwards the Monte Carlo estimate of the tail probability is given by

$$\frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\left(T(\boldsymbol{y}_{[b]}, \boldsymbol{\theta}_{[b]}) \geq T(\boldsymbol{x}, \boldsymbol{\theta}_{[b]})\right).$$

Gelman, Carlin, Stern, and Rubin (2003, p. 164) give an example where the parameter is the mean of the predictive distribution, and the discrepancy measures the difference of the distances of the 10% and 90% data quantiles to that mean. This results in a check for the symmetry fit of the predictive distribution. A similar measure which only includes the model parameters is utilized by Sinharay and Stern (2003, p. 219) to check the normality assumption for the random effects in a hierarchical normal model.

Rather classic regression-diagnostic type checks are presented by Gilks, Richardson, and Spiegelhalter (1998, p. 152). For example, the residual $x - \mathbb{E}(Y)$ or standardized residual $\big(x - \mathbb{E}(Y)\big)/\sqrt{\operatorname{Var}(Y)}$ can be computed for scalar observations $x$. For a set of observations $x_1, \dots, x_n$, the sum of squared standardized residuals gives the $\chi^2$-discrepancy

$$\chi^2 = \sum_{i=1}^{n} \frac{\big(x_i - \mathbb{E}(Y_i)\big)^2}{\operatorname{Var}(Y_i)}.$$

This score depends on the predictive distributions $F_i$ only through the means $\mathbb{E}(Y_i) = \int y \, dF_i(y)$ and the variances $\operatorname{Var}(Y_i) = \int \big(y - \mathbb{E}(Y_i)\big)^2 \, dF_i(y)$. It should be approximately 1 for good predictive performance, so a derived penalty type score is $(\chi^2 - 1)^2$ (Czado, Gneiting, and Held 2009, p. 8). It is interesting that the authors' examples also comprise the PIT, BOT and the conditional predictive ordinate $f_Y(x)$, which is equivalent to the exponent of the logarithmic score.

# 3 Conjugate change point models

In section 3.1, change point models are motivated as a special time series model class. The model framework of general *conjugate* change point models including prior assumptions, posterior inference and handling of missing observations is described in section 3.2. Section 3.3 then proposes approximate sampling schemes for predictive assessment of the one-step-ahead and cross-validation types, which avoid the huge computational effort imposed by the exact sampling schemes. The approaches are contrasted with the goodness-of-fit assessment using posterior-predictive samples. The next three sections are distributions-specific examples of the general framework: While section 3.4 and section 3.5 deal with count likelihoods of Poisson and binomial type, respectively, section 3.6 discusses the appropriate normal model for real-valued time series. The three sections each comprise an extensive case study of data previously analyzed in the literature, and compare the results of exact and approximate predictive assessment. Section 3.7 analyzes a more recent data set of larger dimension, where the exact assessment is not feasible any longer. Finally section 3.8 summarizes the results of this chapter.

## 3.1 Introduction

Change point models for time series assume an (unobserved) partition of the time frame into blocks/segments. In each block, the (unobserved) model parameter is constant. That is, the model parameter seen as a function of the time is a step function, with the steps occurring only at the so-called change points. Conditional on the model parameters, independent observations are recorded. Change point models are special partition models, which also comprise models partitioning higher-dimensional spaces into homogeneous regions. See Hartigan (1990) and Denison, Holmes, Mallick, and Smith (2002, chapter 7) including the references therein for general partition models.

The recorded time can be continuous or discrete. For example, Green (1995, p. 717) analyzes the coal mining data with the points recorded in days over 112 years, using a continuous multiple change points model. In our case study on the same data in section 3.4.2, we use only the year precision, and can thus use our discrete time model. The discrete time case goes back to Chernoff and Zacks (1964). Their normal observations model was later picked up by Yao (1984), who found a more efficient Bayes solution for it.

Barry and Hartigan (1993) also conduct a Bayesian analysis for change point problems, and compare both approaches in a simulation study. They employ MCMC within Gibbs sampling for producing change points draws. The approach implemented in this thesis has been described by Hofmann (2007) for the discrete Poisson-Gamma model. He specialized the approach from Fearnhead (2006, p. 7), who proposed filtering recursions to build a Monte Carlo sampler for the change points samples. This "perfect simulation" avoids the convergence issues inherent to the MCMC solutions.

Usually one is interested in identifying the borders between the blocks, that is one does inference for the change points. Conditional on a change point configuration, the model parameters in the blocks are estimated. Model averaging over multiple change point configurations can be used for marginal inference of the model parameters. The model class has several advantages, with the major one being the adaptive smoothing property: unlike e. g. P-splines, the smoothing effect can be stronger in regions with less variability of the observations and weaker in others with more. We also do not need to directly specify a correlation prior, as it is done with random walk assumptions for the P-spline coefficients. However, to get smoother parameter function estimates, model averaging has to be done, which might be a disadvantage of the model class.

## 3.2 Modelling framework

Section 3.2.1 describes the data situation in which the change point model from section 3.2.2 can be applied. The prior choice in section 3.2.3 ensures that the posterior sampling (section 3.2.4) is easy due to conjugacy of the likelihood and the model parameters prior. Section 3.2.5 discusses necessary changes to the algorithm when some observations are missing.

### 3.2.1 Data

We assume that a time series $\boldsymbol{y} := (y_1, y_2, \dots, y_n)$ of $n$ scalar observations is recorded in the time range $\mathcal{N} = \{1, 2, \dots, n\}$. The index set $\mathcal{N}$ is only used for notational convenience, in reality there will be a (strictly increasing) mapping of indexes to calendar times. In parallel, covariates $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ may be observed. They comprise measurements of variables which are potentially influential for the responses $y_1, y_2, \dots, y_n$. There may also be missing responses $y_t$. But if there are covariates, each observed $y_t$ must have an associated $\boldsymbol{x}_t$ available.

### 3.2.2 Model

At each time point $t \in \mathcal{N}$, a model parameter $\boldsymbol{\xi}_t$ parametrizes the data generating distribution of $y_t$. If the model includes covariates, then also $\boldsymbol{x}_t$ modifies the likelihood $f(y_t \,|\, \boldsymbol{\xi}_t, \boldsymbol{x}_t)$. The dependence of the observations is only generated through the model parameters, and conditional on these and the covariates, the observations are assumed independent:

$$y_t \,|\, \boldsymbol{\xi}_t, \boldsymbol{x}_t \overset{ind}{\sim} f(\cdot \,|\, \boldsymbol{\xi}_t, \boldsymbol{x}_t), \quad t \in \mathcal{N}.$$

For notational brevity we will omit the covariates in the density condition whenever they are not necessary for understanding the formulae.

The characterizing property is the change point model for the parameters $\boldsymbol{\xi}_t$. The number of change points is $k$ and can be any integer between 0 and $n-1$. That means we have $k+1$ unique levels $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \dots, \boldsymbol{\xi}^{(k+1)}$ of the model parameters. The (location of the) change points are

$$\theta_1 < \theta_2 < \cdots < \theta_k \in \{1, 2, \dots, n-1\}.$$

We use the convention that the parameters level changes *after* the change point, so the step function value $\boldsymbol{\xi}_t$ can be written as

$$\boldsymbol{\xi}_t = \sum_{j=1}^{k+1} \mathbb{I}_{(\theta_{j-1}, \theta_j]}(t) \boldsymbol{\xi}^{(j)} \tag{3.2.1}$$

with the start point $\theta_0 := 0$ and the end point $\theta_{k+1} := n$. This means that $\boldsymbol{\xi}_t$ equals the $j$-th level $\boldsymbol{\xi}^{(j)}$ if $t \in (\theta_{j-1}, \theta_j]$. So the $\boldsymbol{\xi}_t$ are determined by the change points parameter $\boldsymbol{\theta} := (\theta_1, \dots, \theta_k)$ and the levels parameter $\boldsymbol{\xi} := (\boldsymbol{\xi}^{(1)}, \dots, \boldsymbol{\xi}^{(k+1)})$; they can be seen as function values $\boldsymbol{\xi}_t(\boldsymbol{\xi}, \boldsymbol{\theta})$ specified by (3.2.1).

### 3.2.3 Prior choice

The prior for the proposed model is naturally split into a prior for the change points, and a prior for the model parameter levels.

#### Prior for the change points

The number $k$ and the locations $\boldsymbol{\theta}$ of the change points are assumed unknown. The change points shall *a priori* follow a Markov process, and a sample path from this process determines the number and the locations of the change points.

The process is specified by the prior transition probabilities

$$\mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1) \tag{3.2.2}$$

which are the probabilities that the $(j+1)$-th change point occurs at time $s = t, t+1, \dots, n$, given the occurrence of the $j$-th change point ($j = 1, 2, \dots, n-1$) at time $t-1$ ($t = j+1, j+2, \dots, n$). Here, the time $s = n$ means that no further change point occurs in the parameter sequence, giving a total of $k = j$ change points. The start probabilities of the Markov process,

$$\mathbb{P}(\theta_1 = s) = \mathbb{P}(\theta_1 = s \,|\, \theta_0 = 0),$$

are generated from the transition probabilities by setting $j = 0$ and $t = 1$, since the start point $\theta_0$ is (by definition) always at time $t - 1 = 0$. Two specific change point priors will be used: the "flat number prior" and the "binomial number prior", which are described in the following.

The "flat number prior" has been used by Hofmann (2007) and Held, Hofmann, Höhle, and Schmid (2006): they place a uniform prior on the number of change points $k$,

$$\mathbb{P}(K = k) = \frac{1}{n}\,\mathbb{I}_{[0,n-1]}(k). \tag{3.2.3}$$

Conditional on the number $k$, they place a uniform prior on the possible configurations $\boldsymbol{\theta}$ with $k$ change points,

$$\mathbb{P}(\boldsymbol{\Theta} = \boldsymbol{\theta} \,|\, K = k) = \binom{n-1}{k}\,\mathbb{I}_{\{\boldsymbol{\theta}\in[1,n-1]^k \,|\, \theta_1<\dots<\theta_k\}}(\boldsymbol{\theta}). \tag{3.2.4}$$

From Hofmann (2007, p. 37) we have that the prior transition probabilities for this prior are

$$\mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1) = \prod_{i=t}^{s-1}\left(1 - \frac{j+1}{i+1}\right)\cdot\left(\frac{j+1}{s+1}\right)^{\mathbb{I}_{[t,n-1]}(s)}. \tag{3.2.5}$$

Note that the factor $(j+1)/(s+1)$ is omitted when $s = n$, because $\mathbb{I}_{[t,n-1]}(n) = 0$. The assumed change point prior implements a model prior with interesting properties. While each dimension $k$ has equal prior probability, the number of models with dimension $k$ increases from $k = 0$ until $k = \lfloor (n-1)/2 \rfloor$ and decreases symmetrically afterwards until $k = n - 1$. So the model with no change points ($k = 0$) has the same prior probability as the model with one level for each observation ($k = n-1$), namely $1/n$. This is the largest prior model probability. By contrast, the models with $k = \lfloor (n-1)/2 \rfloor$ change points have the smallest prior probabilities.

Alternatively, we can use a "binomial number prior" which assigns the event of a change point occurring at a specific time the probability $\pi \in [0, 1]$, identically and independently for all times $t \in \{1, 2, \dots, n-1\}$. So we have $n - 1$ Bernoulli experiments, leading to the binomial distribution $K \sim \mathrm{Bin}(n-1, \pi)$ of the number of change points. Clearly the waiting times between change point times are geometrically distributed, so the prior transition probabilities have the form

$$\mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1) = (1-\pi)^{s-t}\cdot\pi^{\mathbb{I}_{[t,n-1]}(s)}.$$

Therefore Yao (1984, p. 1435) describe the prior as a “discrete renewal process with identically geometrically distributed interarrival times”. This prior was also used by Barry and Hartigan (1993, p. 310) as a special product partition model. Fearnhead (2006, p. 205) generalizes it to the negative binomial distribution.

Note that any valid transition kernels could be used for (3.2.2). For example, Fearnhead (2006, p. 207) places a Poisson prior on the number of change points and draws the locations from an order statistics distribution of uniform draws from the set $\{1, \dots, n-1\}$.

### Prior for the parameters

We specify independent identical prior distributions for the parameter levels $\boldsymbol{\xi}^{(j)}$. These prior distributions have a hyperparameter, say $\boldsymbol{\phi}$, so we assume

$$\boldsymbol{\xi}^{(j)} \overset{iid}{\sim} f(\cdot \,|\, \boldsymbol{\phi}), \quad j = 1, \dots, k+1 \leq n,$$

if the change points configuration $\boldsymbol{\theta}$ is of dimension $k$. For notational brevity, we will omit the hyperparameter $\boldsymbol{\phi}$ from the density condition if it is not essential.

Formally, we could always include $n$ parameter levels in our model, which is the maximum number of possible blocks. However, this is only of theoretical interest, because the unneeded parameter levels would not influence the observations and their posterior distributions would be identical to their prior distributions. Just keep in mind that the parameter levels prior specification is independent of the change points configuration.

The densities $f(\boldsymbol{\xi}^{(j)})$ shall be conjugate to the likelihood $f(y_t \,|\, \boldsymbol{\xi}_t)$. Thus the marginal “block” density for the a parameter block comprising all times in a set $\mathcal{S} \subset \mathcal{N}$ is

$$\begin{aligned} f_{block}(\boldsymbol{y}_{\mathcal{S}}) &:= f(\boldsymbol{y}_{\mathcal{S}} \,|\, y_s, s \in \mathcal{S}, \text{ belong to the same parameter block}) \qquad (3.2.6)\\ &= \int \prod_{t \in \mathcal{S}} f(y_t \,|\, \boldsymbol{\xi}^{(j)}) \cdot f(\boldsymbol{\xi}^{(j)}) \, d\boldsymbol{\xi}^{(j)} \\ &= \frac{\prod_{t \in \mathcal{S}} f(y_t \,|\, \boldsymbol{\xi}^{(j)}) \cdot f(\boldsymbol{\xi}^{(j)})}{f_{block}(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}})} \end{aligned}$$

can be computed analytically, because the block posterior density

$$\begin{aligned} f_{block}(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}) &:= f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}} \text{ and } y_s, s \in \mathcal{S}, \text{ belong to the same parameter block}) \\ & \qquad (3.2.7) \\ &\propto \prod_{t \in \mathcal{S}} f(y_t \,|\, \boldsymbol{\xi}^{(j)}) \cdot f(\boldsymbol{\xi}^{(j)}) \end{aligned}$$

of the parameter level $\boldsymbol{\xi}^{(j)}$ is known. It has the same form as the prior density, $f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{\phi})$, only with an updated hyperparameter, say $\boldsymbol{\phi}_{\mathcal{S}}$, accounting for the new information in the data $\boldsymbol{y}_{\mathcal{S}}$:

$$f_{block}(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}) = f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{\phi}_{\mathcal{S}})$$

The updated hyperparameter can be derived as in the classic case of iid observations from a likelihood which is conjugate to the prior distribution.

So we can calculate the marginal likelihood of a change points configuration $\boldsymbol{\theta}$,

$$\begin{aligned} f(\boldsymbol{y} \,|\, \boldsymbol{\theta}) &= \int f(\boldsymbol{y}, \boldsymbol{\xi} \,|\, \boldsymbol{\theta}) \, d\boldsymbol{\xi} \\ &= \int f(\boldsymbol{y} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}) \, d\boldsymbol{\xi}, \end{aligned}$$

by noting that both the likelihood

$$f(\boldsymbol{y} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{j=1}^{k+1} \prod_{t \in (\theta_{j-1}, \theta_j]} f(y_t \,|\, \boldsymbol{\xi}^{(j)})$$

and the conditional prior of $\boldsymbol{\xi}$

$$f(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{k+1} f(\boldsymbol{\xi}^{(j)})$$

factorize into the $k+1$ blocks because of our independence assumptions: The integral becomes

$$f(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \prod_{j=1}^{k+1} f_{block}(\boldsymbol{y}_{(\theta_{j-1}, \theta_j]}). \tag{3.2.8}$$

The structure of the described model is summarized in Figure 3.1: The observations $y_t$ are conditionally independent of each other, given the covariates $\boldsymbol{x}_t$ and the parameters $\boldsymbol{\xi}_t$, which specify the data generating distribution from a given likelihood family. The change points $\boldsymbol{\theta}$ and the parameters levels $\boldsymbol{\xi}$ determine the parameters $\boldsymbol{\xi}_t$. The parameter levels $\boldsymbol{\xi}$ have prior parameters $\boldsymbol{\phi}$ specifying the form of the conjugate prior distribution. The prior distribution of the configuration $\boldsymbol{\theta}$ can be arbitrarily defined through the prior transition probabilities (3.2.2) and is not shown in Figure 3.1.

### 3.2.4 Posterior

The advantage of the conjugacy is that the efficient forward-backward algorithm described by Hofmann (2007) can be used to directly sample from the marginal posterior $f(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ of the change points. Sampling from the conditional posterior $f(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}, \boldsymbol{y})$ of the levels parameter $\boldsymbol{\xi}$ is also easy due to the choice of the conjugate parameter prior. Thus, in order to estimate the full posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, ordinary Monte Carlo estimation is possible.

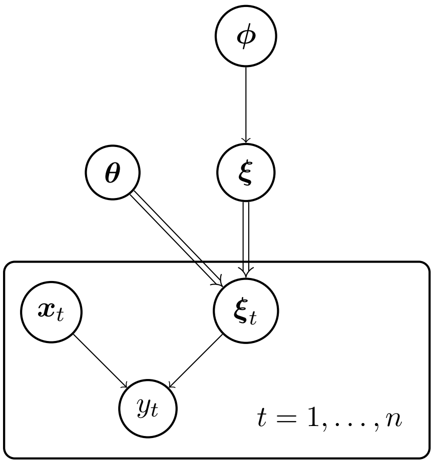


***Figure 3.1*** – *Graphical model of the proposed framework.*

### Sampling the change points

In the so-called forward step, we can compute the following recursion for the conditional density of the counts $y_t, y_{t+1}, \ldots, y_n$ ($t \in \{2, 3, \ldots, n-1\}$) given that the $j$-th change point ($j \in \{1, 2, \ldots, t-1\}$) occurred just before:

$$\begin{aligned} f(\boldsymbol{y}_{[t,n]} \,|\, \theta_j = t-1) &= \sum_{s=t}^{n} f(\boldsymbol{y}_{[t,n]}, \theta_{j+1} = s \,|\, \theta_j = t-1) \\ &= \sum_{s=t}^{n} f(\boldsymbol{y}_{[t,n]} \,|\, \theta_{j+1} = s, \theta_j = t-1)\, \mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1) \\ &= \sum_{s=t}^{n} f_{block}(\boldsymbol{y}_{[t,s]}) f(\boldsymbol{y}_{[s+1,n]} \,|\, \theta_{j+1} = s)\, \mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1). \end{aligned} \tag{3.2.9}$$

Note that for ease of notation the terms $f(\boldsymbol{y}_{[n+1,n]} \,|\, \theta_{j+1} = n)$ shall evaluate to unity for all $j$, similarly as an empty product from $n+1$ to $n$.

So the conditional densities $f(\boldsymbol{y}_{[t,n]} \,|\, \theta_j = t-1)$ conditioning on the $j$-th change point position, which are indexed by the time $t$, depend on the densities

$$f(\boldsymbol{y}_{[t+1,n]} \,|\, \theta_{j+1} = t), \ldots, f(\boldsymbol{y}_{[n,n]} \,|\, \theta_{j+1} = n-1), f(\boldsymbol{y}_{[n+1,n]} \,|\, \theta_{j+1} = n) \equiv 1$$

for the $(j+1)$-th change point. The start for this recursion is the single density value at $j = n-1$ with $t = n$,

$$f(\boldsymbol{y}_{[n,n]} \,|\, \theta_{n-1} = n-1) = f_{block}(y_n),$$

because the probability $\mathbb{P}(\theta_n = n \,|\, \theta_{n-1} = n-1)$ equals unity. Afterwards, the conditional densities for $j = n-2, n-3, \ldots, 1$ can be computed. Finally, by setting $j = 0$ and $t = 1$

the unconditional density of the whole time series $\boldsymbol{y}$ is obtained, which is the marginal likelihood of our whole model:

$$f(\boldsymbol{y}) = \sum_{s=1}^{n} f_{block}(\boldsymbol{y}_{[1,s]}) f(\boldsymbol{y}_{[s+1,n]} \,|\, \theta_1 = s)\, \mathbb{P}(\theta_1 = s). \tag{3.2.10}$$

The backward step consists of computing the posterior transition probabilities of the change point locations, using the conditional densities from the forward step. From Hofmann (2007, p. 38), we have

$$\mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1, \boldsymbol{y}) = \frac{f_{block}(\boldsymbol{y}_{[t,s]}) f(\boldsymbol{y}_{[s+1,n]} \,|\, \theta_{j+1} = s)\, \mathbb{P}(\theta_{j+1} = s \,|\, \theta_j = t-1)}{f(\boldsymbol{y}_{[t,n]} \,|\, \theta_j = t-1)}, \tag{3.2.11}$$

for next change point times $s = t, t+1, \dots, n$, for last change point times $t = j+1, j+2, \dots, n$ and for last change point indexes $j = 1, 2, \dots, n-1$. Similarly as for the prior Markov process of the change points, the posterior starting distribution is obtained with $j = 0$, $t = 1$, yielding

$$\mathbb{P}(\theta_1 = s \,|\, \boldsymbol{y}) = \frac{f_{block}(\boldsymbol{y}_{[1,s]}) f(\boldsymbol{y}_{[s+1,n]} \,|\, \theta_1 = s)\, \mathbb{P}(\theta_1 = s)}{f(\boldsymbol{y})}, \tag{3.2.12}$$

for the locations $s = 1, 2, \dots, n$ of the first change point $\theta_1$.

In order to sample one posterior realization of the change point configuration $\boldsymbol{\theta}$, first draw the first change point location $\theta_1$ from the posterior starting distribution in (3.2.12) and set $j = 1$. Second, if the $j$-th change point is at $t - 1 < n$, then draw the location of the next change point $\theta_{j+1}$ from the transition distribution in (3.2.11). Afterwards increment $j$ and repeat the second step. However, if $\theta_j = n$, all $k := j - 1$ change points for the sample are already there, and the posterior realization is $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. This may be an empty tuple if $k = 0$ and there are no change points at all.

In the sequential sampling algorithm, the posterior probability of a change points configuration can be computed sequentially as well. If the first change point is at $s$, we initialize the value with $\mathbb{P}(\theta_1 = s)$ from (3.2.12). Note that this probability has already been computed in order to sample the first change point location. Afterwards, until the change point location $n$ is drawn (what finishes the configuration sample), multiply the saved value with the appropriate transition probability from (3.2.11). Again, this probability is available anyway, so there is no relevant overhead from computing the posterior probability $\mathbb{P}(\boldsymbol{\Theta} = \boldsymbol{\theta} \,|\, \boldsymbol{y})$ of a sample $\boldsymbol{\theta}$. These probabilities can then later be used to identify the *maximum a posteriori* (MAP) change point sample $\boldsymbol{\theta}_{MAP}$ with the highest posterior probability. If one is interested in a single step function for the description of the data $\boldsymbol{y}$, then this model $\boldsymbol{\theta}_{MAP}$ is "the best" in terms of the used probabilistic modelling of the data process which has been found in the sampling process.

### Sampling the parameters given the change points

Conditional on the partition $\boldsymbol{\theta}$ with $k$ change points, the posterior density of the levels $\boldsymbol{\xi}$ is

$$\begin{aligned} f(\boldsymbol{\xi} \mid \boldsymbol{\theta}, \boldsymbol{y}) &\propto f(\boldsymbol{y} \mid \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi} \mid \boldsymbol{\theta}) \\ &= \prod_{j=1}^{k+1} \prod_{t \in (\theta_{j-1}, \theta_j]} f(y_t \mid \boldsymbol{\xi}^{(j)}) f(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{\phi}) \\ &\propto \prod_{j=1}^{k+1} f_{block}(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{y}_{(\theta_{j-1}, \theta_j]}, \boldsymbol{\phi}) \\ &= \prod_{j=1}^{k+1} f(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{\phi}_{(\theta_{j-1}, \theta_j]}). \end{aligned}$$

So we can sample from the $k+1$ independent block posterior distributions $f(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{\phi}_{(\theta_{j-1}, \theta_j]})$, which are updated prior distributions with parameters $\boldsymbol{\phi}_{(\theta_{j-1}, \theta_j]}$, to obtain the realizations of the parameter levels $\boldsymbol{\xi}^{(j)}$. Together with $\boldsymbol{\theta}$, these yield the samples of the parameters $\boldsymbol{\xi}_t$ via the deterministic link (3.2.1).

At last, we have produced a posterior sample of the parameters trend $\{\boldsymbol{\xi}_t\}_{t=1}^{n}$. Note that only this trend can be sensibly compared between different change point samples $\boldsymbol{\theta}$, but not the unique levels $\boldsymbol{\xi}^{(j)}$. For example, we can use Bayesian model averaging of the different step functions if we are only interested in the trend and not in the change points. This is very easy because the samples $\boldsymbol{\xi}_{t,[b]}$, $b = 1, \dots, B$, say, are samples from the marginal posterior

$$f(\boldsymbol{\xi}_t \mid \boldsymbol{y}) = \int f(\boldsymbol{\xi}_t \mid \boldsymbol{\theta}, \boldsymbol{y}) f(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}.$$

So if we just “forget” the change points configuration samples $\boldsymbol{\theta}_{[b]}$, we have averaged the model parameters over the change point models. While this model averaged trend will still be a step function, it will typically be smoother and is able to capture big jumps better than e. g. splines based approaches with global smoothness assumptions.

### Computational considerations

Usually at least a few thousand samples will be required for serious posterior inference. Then, it is advisable to move on to the next change points in parallel across the samples, finishing those samples where $\theta_j = n$ is generated. So if there are $m$ samples with $j$-th change point at $t-1$, then we need to compute the posterior probabilities $\mathbb{P}(\theta_{j+1} = s \mid \theta_j = t-1, \boldsymbol{y})$ only once and not $m$ times.

In the practical implementation of the posterior transition probabilities computation, a trade-off between faster sampling and less memory consumption must be made. On the one hand, saving all transition probabilities from (3.2.11) before the start would accelerate the sampling, because we would not need to compute them again during sampling. In fact, not only the denominators but also the numerators in (3.2.11) are all computed in the forward step: they are summed up in (3.2.9). But the memory consumption for these numerators is cubic in the length $n$ of the time series: Precisely, there are

$$\sum_{j=1}^{n-1} \sum_{t=j+1}^{n} \sum_{s=t}^{n} 1 = \frac{n^3 - n}{6}$$

numerator values. On the other hand, computing the necessary numerators again during the sampling needs more CPU time (how much more depends on the homogeneity of the samples), but saves us memory. And the memory consumption can be high: For $n = 400$, we need to save 10 666 600 numbers with double precision, giving a memory consumption of approximately 85.33 MB, because in the used GCC implementation 64 bits are used for a double number. This is still manageable on recent personal computers. But already for a ten times longer time series, the memory usage is almost 1000 times larger and cannot be handled easily any longer. Therefore in the implementation, for $n \leq 400$ the faster approach is used, while for $n > 400$, the more parsimonious strategy is applied.

Of course, (only) in principle the sampling scheme for the change points configurations is not necessary: The posterior probability of every $\boldsymbol{\theta}$ can be computed via Bayes' theorem

$$\mathbb{P}(\boldsymbol{\Theta} = \boldsymbol{\theta} \,|\, \boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{\Theta} = \boldsymbol{\theta}) f(\boldsymbol{y} \,|\, \boldsymbol{\theta})}{f(\boldsymbol{y})},$$

because the prior probability is determined by (3.2.3) and (3.2.4), and the likelihood (3.2.8) as well as the normalization constant (3.2.10) can be computed. But the huge dimension of the model space renders the use of a general purpose sampler for finite discrete distributions infeasible. There are $2^{n-1}$ possible change point configurations, so even for a moderate length $n = 100$ there are approximately $6.34 \cdot 10^{29}$ possibilities.

### 3.2.5 Handling of missing data

Let $o_t \in \{0, 1\}$ be the observation indicator for the response value $y_t$. So we assume the data is available as the length $n$ vectors $\boldsymbol{y}$, $\boldsymbol{o}$ and possibly $\boldsymbol{x}$, where for $o_t = 0$ the response is missing and the saved value $y_t$ is just a dummy which will not be processed. The times for which the responses are missing are collected in the set $\mathcal{M} := \{t \,|\, o_t = 0\}$, while the times with observed responses form the complementary set $\mathcal{N} \setminus \mathcal{M}$. What modifications are necessary to accommodate the case when $\mathcal{M} \neq \emptyset$, and we want to sample from $f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N} \setminus \mathcal{M}})$?

Algorithmically simplest, but computationally demanding, would be Gibbs sampling with $\boldsymbol{y}_{\mathcal{M}}$ as an auxiliary variable, whose current sample $\boldsymbol{y}^*_{\mathcal{M}}$ is initialized at some valid point inside the support of the observation density before the loop. If there are also missing covariates $\boldsymbol{x}_t$ for $o_t = 0$, then similar values like those available can be imputed into the likelihood. The MCMC algorithm would iterate between two steps:

1. In the first step, $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are drawn from the full posterior $f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}^*_{\mathcal{M}}, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{M}})$, using the sampling scheme exactly as in section 3.2.4. Thus, we draw from the full conditional distribution of $(\boldsymbol{\xi}, \boldsymbol{\theta})$.

2. In the second step, the auxiliary variable $\boldsymbol{y}^*_{\mathcal{M}}$ is drawn from the likelihood $f(\boldsymbol{y}_{\mathcal{M}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta})$, that is $|\mathcal{M}|$ independent samples with the parameters being determined by the current samples of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are generated:

$$y^*_t \overset{ind}{\sim} f(\cdot \,|\, \boldsymbol{\xi}_t, \boldsymbol{x}_t), \quad t \in \mathcal{M}.$$

   This distribution is the full conditional distribution of $y_t$, resulting in a draw from the full conditional distribution of $\boldsymbol{y}_{\mathcal{M}}$.

The Markov chain then eventually converges to $f(\boldsymbol{y}_{\mathcal{M}}, \boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{M}})$. If we are not interested in the distribution of $\boldsymbol{y}_{\mathcal{M}}$, we can just use the samples from the marginal distribution $f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{M}})$. A major disadvantage of this Gibbs sampler is that the forward step has to be recomputed in each iteration, because the auxiliary variable sample $\boldsymbol{y}^*_{\mathcal{M}}$ changes. Furthermore, convergence diagnosis for the Markov chain must be done. These two issues render the Gibbs sampling approach unusable for all practical purposes.

Much better is drawing the parameter levels and the change points directly from $f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{M}})$. This requires only one forward step, and is therefore implemented in the supplementary `R`-package. The idea is that the definition of the times $1, 2, \dots, n$ does not change, but that those in $\mathcal{M}$ do not have associated observed responses. So it will be possible to have a change point $\theta_j = t$ at a missing time $t \in \mathcal{M}$. The necessary modification is to replace $\boldsymbol{y}$ with $\boldsymbol{y}_{\mathcal{N}\setminus\mathcal{M}}$ everywhere, because we want to condition on the observed data only. This means that all sets of the form $\boldsymbol{y}_{[t,s]}$ are replaced with $\boldsymbol{y}_{[t,s]\setminus\mathcal{M}}$.

In the forward-backward algorithm, all conditional densities and the resulting transition probabilities derive from the block marginal likelihoods defined in (3.2.6). For the recursion in (3.2.9), functions

$$g_t(s) := f_{block}(\boldsymbol{y}_{[t,s]\setminus\mathcal{M}}) = \frac{\prod_{r\in[t,s]\setminus\mathcal{M}} f(y_r \,|\, \boldsymbol{\xi}^{(j)}) \cdot f(\boldsymbol{\xi}^{(j)})}{f_{block}(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{y}_{[t,s]\setminus\mathcal{M}})}$$

must be evaluated at $s = t, t+1, \dots, n$, for decreasing start times $t$. Obviously the set $[t,s] \setminus \mathcal{M}$ only changes at times $s \notin \mathcal{M}$, so $g_t(s)$ is a step function with jumps at $s \notin \mathcal{M}$. If already the first response $y_t$ in the window $[t,n]$ is missing ($s = t \in \mathcal{M}$), then the first function value is

$$g_t(t) = f_{block}(\boldsymbol{y}_\emptyset) = \frac{\prod_{r \in \emptyset} f(y_r \mid \boldsymbol{\xi}^{(j)}) \cdot f(\boldsymbol{\xi}^{(j)})}{f_{block}(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{y}_\emptyset)} = 1,$$

because the empty product evaluates to 1 and $f_{block}(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{y}_\emptyset) \equiv f(\boldsymbol{\xi}^{(j)})$. Using this simple modification, we can sample from $f(\boldsymbol{\theta} \mid \boldsymbol{y}_{\mathcal{N} \setminus \mathcal{M}})$ using the otherwise unchanged forward-backward algorithm.

Similarly, to draw the $j$-th parameter level $\boldsymbol{\xi}^{(j)}$ for a given change point configuration $\boldsymbol{\theta}$, we use the updated parameter level density $f(\boldsymbol{\xi}^{(j)} \mid \boldsymbol{\phi}_{(\theta_{j-1},\theta_j] \setminus \mathcal{M}})$. Note that if there are no observations from $\boldsymbol{y}_{\mathcal{N} \setminus \mathcal{M}}$ in the $j$-th block, then we draw $\boldsymbol{\xi}^{(j)}$ from its prior distribution. So we really need a proper prior for the parameter levels, because otherwise we could not sample from it.

## 3.3 Exact and approximate predictive assessment

This section introduces the key topic of this chapter: the comparison of five different predictive sampling schemes. Section 3.3.1 describes the exact and approximate sampling schemes for the one-step-ahead prediction, while section 3.3.2 does the same for the general cross-validated prediction. Both sections discuss how logarithmic scores can be computed analytically for the model class of this chapter. Section 3.3.3 finally turns to the posterior-predictive sampling, which can be useful for goodness-of-fit assessment of the change point models. The last section 3.3.4 compares the sampling schemes definitions and summarizes the logarithmic score estimation results.

### 3.3.1 One-step-ahead predictive assessment

Given the time series from time 1 to time $n$, how well can our model predict the observation at the next time $n + 1$? That is, how good are the one-step-ahead predictions in our change point model? This classic task of time series models has been called “prequential forecasting” by Dawid (1984, p. 278), merging the adjectives of the terms *pr*obability forecasting and s*equential* prediction.

#### Exact sampling

One way to assess the one-step-ahead prediction performance in our model is to try the prediction for the counts at times $t + 1 = 1, 2, \dots, n$, if we feed our algorithm only with the counts at times $1, 2, \dots, t$. For $t = 0$, we predict the first observation $y_1$ from the prior predictive distribution $f(y_1)$. The forward and backward steps must thus be computed $n - 1$ times, one time less than the number of observations because the prior predictive samples are directly obtained from the data generating distribution parametrized by prior parameter samples. The following sampling details are tailored to the more complicated case $t > 1$. Note that a particle filter algorithm could avoid the repeated forward and backward computations implied by our posterior sampling approach from section 3.2.4 (see Doucet, De Freitas, and Gordon (2001) for sequential Monte Carlo methods).

For the sampling from the predictive density $f(y_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$, the idea is to use the conditional independence of the observations $y_t$ given the parameters $\boldsymbol{\xi}_t$:

$$f(y_{t+1} \,|\, \boldsymbol{y}_{[1,t]}) = \int f(y_{t+1} \,|\, \boldsymbol{\xi}_{t+1}) f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{y}_{[1,t]}) \, d\boldsymbol{\xi}_{t+1}$$

So if we can sample from $f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$, we just plug the realization $\boldsymbol{\xi}_{t+1}$ into the likelihood and keep $y^*_{t+1} \sim f(\cdot \,|\, \boldsymbol{\xi}_{t+1}, \boldsymbol{x}_{t+1})$ as a sample from the predictive distribution.

One solution is to give the algorithm the response and observation indicator vectors $(y_1, \dots, y_t, 0)$ and $(o_1, \dots, o_t, 0)$ to mark $y_{t+1}$ as missing. This naturally produces samples from $f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$.

Another solution exploits the sequential structure of the model. Consider

$$f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{y}_{[1,t]}) = \iint f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{\xi}_t, \boldsymbol{\theta}_{[1,t]}) f(\boldsymbol{\xi}_t, \boldsymbol{\theta}_{[1,t]} \,|\, \boldsymbol{y}_{[1,t]}) \, d\boldsymbol{\xi}_t \, d\boldsymbol{\theta}_{[1,t]},$$

where $\boldsymbol{\theta}_{[1,t]}$ is the change points configuration in the time series $\boldsymbol{y}_{[1,t]}$: The next parameter $\boldsymbol{\xi}_{t+1}$ only depends on the last parameter $\boldsymbol{\xi}_t$ and $\boldsymbol{\theta}_{[1,t]}$, because if there is no change point at time $t$, then $\boldsymbol{\xi}_{t+1} \equiv \boldsymbol{\xi}_t$, else we draw the next parameter from the prior $f(\boldsymbol{\xi} \,|\, \boldsymbol{\phi})$. This mixture of a point mass at $\boldsymbol{\xi}_t$ and the prior distribution is weighted by the probability of a change point at time $t$ given $\boldsymbol{\theta}_{[1,t]}$. If $\boldsymbol{\theta}_{[1,t]}$ contains $k$ change points with the last change point occurring at time $s$, this probability is

$$\mathbb{P}(\theta_{k+1} = t \,|\, \theta_k = s, \boldsymbol{y}_{[1,t]}) = \mathbb{P}(\theta_{k+1} = t \,|\, \theta_k = s),$$

equal to the respective prior transition probability. The reason is that the event of a change point occurrence at time $t$ is independent of the observations until time $t$ – they are happening before this change point could have an effect. For the flat number prior, this probability equals $(k + 1)/(t + 1)$, which is the "success probability" already visible in (3.2.5). See Hofmann (2007, p. 34) or Held, Hofmann, Höhle, and Schmid (2006, section 2.6) for more details on the derivation. For the binomial number prior, this probability equals the hyperparameter $\pi$. So if we have sampled $\boldsymbol{\xi}_t$ and $\boldsymbol{\theta}_{[1,t]}$ with $k$ change points, afterwards we sample

$$\boldsymbol{\xi}_{t+1} \sim \{1 - \mathbb{P}(\theta_{k+1} = t \,|\, \theta_k = s)\} \, \delta_{\boldsymbol{\xi}_t} + \mathbb{P}(\theta_{k+1} = t \,|\, \theta_k = s) f(\boldsymbol{\xi} \,|\, \boldsymbol{\phi}).$$

Both solutions require $n - 1$ forward steps: for the first solution, the algorithm must be run with the observation indicator vectors

$$(o_1, 0, \dots, 0), (o_1, o_2, 0, \dots, 0), \dots, (o_1, o_2, \dots, o_{n-1}, 0),$$

and for the second solution we must start the algorithm for each of the data vectors

$$y_1, \boldsymbol{y}_{[1,2]}, \dots, \boldsymbol{y}_{[1,n-1]}.$$

Note that this second solution is the basis for the approximate sampling described below.

### Approximate sampling

The exact sampling from $f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$ is computationally expensive, because for every time $t$, a new forward step is necessary. An approximate version of the sampling is inspired by Marshall and Spiegelhalter (2003).

The idea is to use the whole time series $\boldsymbol{y}$ for only one forward-backward run. That is, the same sampling probabilities (3.2.12) and (3.2.11) are used for *all* learning sets $\{\boldsymbol{y}_{[1,t]}\}$, $t = 1, 2, \dots, n-1$. So the change point locations $\boldsymbol{\theta}_{[1,t]}$ up to time $t$ which are used above are not drawn from $f(\boldsymbol{\theta}_{[1,t]} \,|\, \boldsymbol{y}_{[1,t]})$ but from $f(\boldsymbol{\theta}_{[1,t]} \,|\, \boldsymbol{y})$. In practice, the change points sampler is run on the whole data $\boldsymbol{y}$ and produces samples of $\boldsymbol{\theta}$. Then for the processing of the learning set $\boldsymbol{y}_{[1,t]}$, all change points at times $t, t+1, \dots, n-1$ are deleted from the sample vectors to produces approximate samples of $\boldsymbol{\theta}_{[1,t]}$, which conform with the maximum change point time $t-1$ for the reduced data $\boldsymbol{y}_{[1,t]}$. (The conventional "change point" at the last time $t$ is deterministic and is of course included.)

However, sampling the parameter level $\boldsymbol{\xi}^{(k+1)} \equiv \boldsymbol{\xi}_t$ of the block including the time $t$ from the correct density $f(\boldsymbol{\xi}^{(k+1)} \,|\, \boldsymbol{\theta}_{[1,t]}, \boldsymbol{y}_{[1,t]})$ is easy: it is just the block posterior distribution $f_{block}(\boldsymbol{\xi}^{(k+1)} \,|\, \boldsymbol{y}_{(\theta_k,t]}, \boldsymbol{\phi})$. Therefore, the unknown part $\boldsymbol{y}_{[t+1,n]}$ influences the parameter sample $\boldsymbol{\xi}_t$ only indirectly, through the "pruned" samples $\boldsymbol{\theta}_{[1,t]}$ obtained from $\boldsymbol{\theta}$. Hence, the conservatism concerning the sampling of $\boldsymbol{\xi}_{t+1}$ and $y^*_{t+1}$, which is the price for the reduction of computing time in this approximate sampling scheme, should be moderate.

The proposed approximation of the one-step-ahead predictive density is thus

$$\begin{aligned}
f(y_{t+1} \,|\, \boldsymbol{y}_{[1,t]}) &= \iiint f(y_{t+1} \,|\, \boldsymbol{\xi}_{t+1}) f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{\xi}_t, \boldsymbol{\theta}_{[1,t]}) f(\boldsymbol{\xi}_t, \boldsymbol{\theta}_{[1,t]} \,|\, \boldsymbol{y}_{[1,t]}) \, d\boldsymbol{\xi}_{t+1} \, d\boldsymbol{\xi}_t \, d\boldsymbol{\theta}_{[1,t]} \\
&= \iiint f(y_{t+1} \,|\, \boldsymbol{\xi}_{t+1}) f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{\xi}_t, \boldsymbol{\theta}_{[1,t]}) f(\boldsymbol{\xi}_t \,|\, \boldsymbol{\theta}_{[1,t]}, \boldsymbol{y}_{[1,t]}) f(\boldsymbol{\theta}_{[1,t]} \,|\, \boldsymbol{y}_{[1,t]}) \, d\boldsymbol{\xi}_{t+1} \, d\boldsymbol{\xi}_t \, d\boldsymbol{\theta}_{[1,t]} \\
&\approx \iiint f(y_{t+1} \,|\, \boldsymbol{\xi}_{t+1}) f(\boldsymbol{\xi}_{t+1} \,|\, \boldsymbol{\xi}_t, \boldsymbol{\theta}_{[1,t]}) f(\boldsymbol{\xi}_t \,|\, \boldsymbol{\theta}_{[1,t]}, \boldsymbol{y}_{[1,t]}) f(\boldsymbol{\theta}_{[1,t]} \,|\, \boldsymbol{y}) \, d\boldsymbol{\xi}_{t+1} \, d\boldsymbol{\xi}_t \, d\boldsymbol{\theta}_{[1,t]} \\
&=: \tilde{f}(y_{t+1} \,|\, \boldsymbol{y}_{[1,t]}).
\end{aligned}$$

### Analytical logarithmic scores

The proposed conjugate change point model allows the analytical computation of the one-step-ahead logarithmic scores

$$-\log f(y_1), -\log f(y_2 \,|\, y_1), \dots, -\log f(y_n \,|\, \boldsymbol{y}_{[1,n-1]}),$$

because all marginal likelihoods $f(\boldsymbol{y}_{[1,t]})$, $t = 1, \dots, n$ can be calculated in an exact one-step-ahead sampling loop: While for $t = 1$, we just use $f(y_1) = f_{block}(y_1)$ and already have the first score, for $t \geq 2$ we obtain $f(\boldsymbol{y}_{[1,t]})$ from (3.2.10) as a by-product from the (reduced) posterior sampling given the data $\boldsymbol{y}_{[1,t]}$. Remember that the marginal likelihood is computed at the end of the forward step, which is necessary for the change points sampling in the backward step. Having finished the exact one-step-ahead validation loop, we can compute the remaining logarithmic scores

$$-\log f(y_{t+1} \,|\, \boldsymbol{y}_{[1,t]}) = \log f(\boldsymbol{y}_{[1,t]}) - \log f(\boldsymbol{y}_{[1,t+1]}), \quad t = 1, \dots, n-1.$$

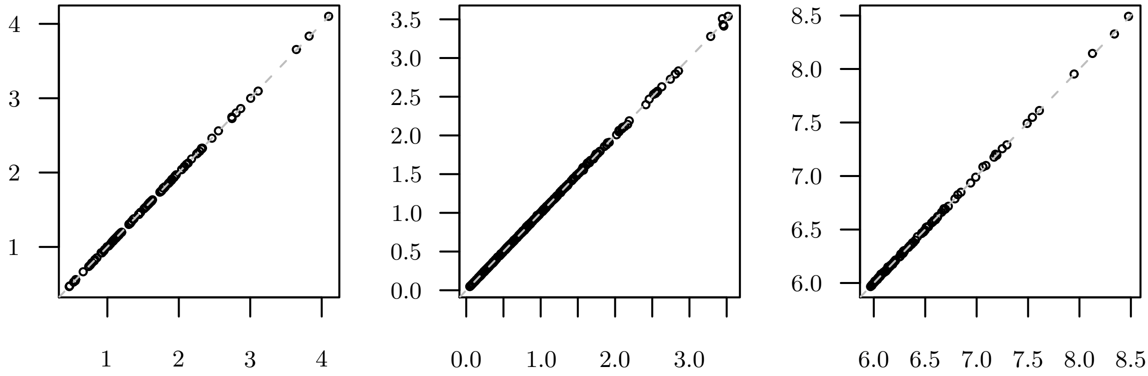
Note that the sum of the one-step-ahead logarithmic scores equals the negative marginal log-likelihood. The well-known decomposition of the marginal likelihood of the vector $\boldsymbol{y}$ into the conditional scalar densities,

$$f(\boldsymbol{y}) = f(y_n \,|\, \boldsymbol{y}_{[1,n-1]}) f(y_{n-1} \,|\, \boldsymbol{y}_{[1,n-2]}) \cdots f(y_2 \,|\, y_1) f(y_1),$$

is the equivalent on the multiplicative scale. Therefore the mean one-step-ahead log-score can be computed directly from the marginal likelihood in the conjugate change point model as $\overline{LogS} = -\frac{1}{n} \log f(\boldsymbol{y})$. Since this is a strictly monotone transformation, the model comparison based on the one-step-ahead log-score is equivalent to that based on the marginal likelihood. Another consequence is that the mean one-step-ahead log-score of the reversed time series is identical to the log-score of the original time series, because the assumed prior and likelihood are invariant to the time direction and so the marginal likelihood is identical.

The estimation of the logarithmic scores using the Monte Carlo approach on page 14 is nevertheless sensible, because we can thus assess the Monte Carlo error which also contributes to the difference between exact sampling and approximate sampling results. We have compared the analytical one-step-ahead log-scores with the corresponding exact sampling log-scores for all examined models in the case studies from sections 3.4.2, 3.5.2 and 3.6.2. The maximum absolute differences for the three sections were 0.051, 0.065 and 0.1, while the mean deviances were 0.004, 0.003 and 0.005, respectively. These Monte Carlo errors are very small compared to the approximate sampling errors, with maximum deviances 1.924, 1.705 and 1.253, and mean deviances 0.107, 0.098 and 0.079, respectively in the three sections when all models are pooled. For illustration we show comparison plots of exact sampling versus analytical log-scores for three selected models in Figure 3.2. Only in panel (b) some points in the upper right corner are lying slightly away from the identity line.

***Figure 3.2*** – *Comparison of analytical one-step-ahead log-scores (x-axis) and corresponding exact sampling log-scores (y-axis), for three models from sections 3.4.2, 3.5.2 and 3.6.2.*



(a) Coal-mining data model 1 (b) Tokyo rainfall data model 2 (c) Nile discharge data model 3

### 3.3.2 Cross-validation assessment

How well can our model predict the observations $y_t$ at times $t \in \mathcal{T} \subset \{1, 2, \dots, n\}$ (the test set), if we only provide it with the observations in $\mathcal{N} \setminus \mathcal{T}$ (the learning set)?

To answer this question, usually cross-validation is done, that is, the original data $\boldsymbol{y}$ is repeatedly split into disjoint test and learning sets. A popular choice is to define the test set $\mathcal{T} := \{t\}$ for all times $t = 1, 2, \dots, n$ in turn. This corresponds to a leave-one-out cross-validation of the model. Yet other choices may make sense, e. g. leave out whole months in turn if the time resolution is one day. Also, the prediction of the remaining times in $\mathcal{T} := \{t + 1, t + 2, \dots, n\}$ can be of interest. If $t < n - 1$, this would correspond to a multiple-steps-ahead prediction because $|\mathcal{T}| > 1$, in contrast to the one-step-ahead prediction being assessed in section 3.3.1.

Again we want to base the predictive assessment on samples $\boldsymbol{y}^*_{\mathcal{T}}$ from the predictive density $f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$. So how can we efficiently generate such samples?

#### Exact sampling

For exact sampling from $f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$, we will use the conditional independence of the observations given the parameters and change points. This leads to

$$f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}) = \iint f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})\, d\boldsymbol{\xi}\, d\boldsymbol{\theta},$$

meaning that sampling from the posterior density $f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$ based on the learning set, followed by sampling from the likelihood $f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta})$, produces the required samples $\boldsymbol{y}^*_{\mathcal{T}}$.

The sampling from the learning set posterior is easy after the discussion of missing data handling in section 3.2.5: If the original data comprised the indicator vector $\boldsymbol{o}$, we just mark each time in the test set $\mathcal{T}$ as missing to form the learning set $\mathcal{N} \setminus \mathcal{T}$. So we let our sampler run with indicator variables $\tilde{o}_t := o_t \cdot (1 - \mathbb{I}_{\mathcal{T}}(t))$. This careful handling is necessary because if the observation $y_t$ was missing in the original data ($o_t = 0$), it is of course also missing in the learning set. Finally, we get samples from $f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$.

### Approximate sampling

The exact sampling from the cross-validation density $f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$ requires much computational effort, because for every test set $\mathcal{T}$, a new forward step is necessary to be able to draw from $f(\boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$ in the backward step. This is relevant because usually the number of test sets increases at the order $O(n)$, where $n$ is the number of time points. As the effort for a forward step is $O(n^3)$, the cross-validation effort is usually $O(n^4)$.

An approximate version does only once sample from $f(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ and thus requires just one forward step for all cross-validation iterations. Given the change points $\boldsymbol{\theta}$, this "Marshall-Spiegelhalter version" proceeds with sampling $\boldsymbol{\xi} \,|\, \boldsymbol{\theta}, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}$ from the correct conditional posterior distributions, as was sketched at the end of section 3.2.5. The final generation of the predictive samples $\boldsymbol{y}^*_{\mathcal{T}}$ by sampling from the likelihood $f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta})$ remains unchanged. That means, we make the approximation

$$\begin{aligned} f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}) &= \iint f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi}, \boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}) \, d\boldsymbol{\xi} \, d\boldsymbol{\theta} \\ &= \iint f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}) f(\boldsymbol{\theta} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}) \, d\boldsymbol{\xi} \, d\boldsymbol{\theta} \\ &\approx \iint f(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi} \,|\, \boldsymbol{\theta}, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}) f(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \, d\boldsymbol{\xi} \, d\boldsymbol{\theta} \\ &=: \tilde{f}(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}}). \end{aligned}$$

### Analytical logarithmic scores

The proposed conjugate change point model allows the analytical computation of the leave-one-out logarithmic scores

$$-\log f(y_1 \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\{1\}}), -\log f(y_2 \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\{2\}}), \ldots, -\log f(y_n \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\{n\}}),$$

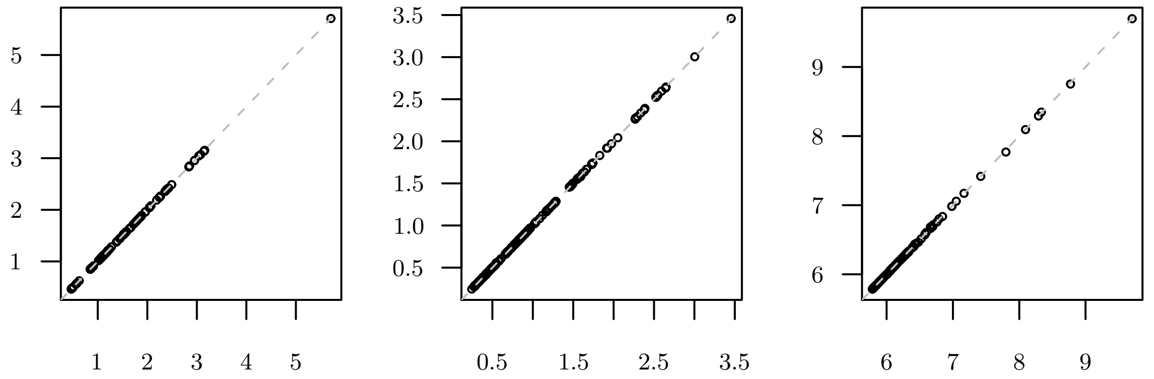
because all marginal likelihoods $f(\boldsymbol{y}_{\mathcal{N}\setminus\{t\}})$, $t = 1, \ldots, n$ can be calculated from (3.2.10) and are a by-product of the reduced exact posterior sampling given the data $\boldsymbol{y}_{\mathcal{N}\setminus\{t\}}$. Having finished the exact leave-one-out validation loop, we can compute the logarithmic scores as

$$-\log f(y_t \,|\, \boldsymbol{y}_{\mathcal{N}\setminus\{t\}}) = \log f(\boldsymbol{y}_{\mathcal{N}\setminus\{t\}}) - \log f(\boldsymbol{y}), \quad t = 1, \ldots, n,$$

using the full data marginal likelihood $f(\boldsymbol{y})$ from the full data forward step. Analogously, more general cross-validation logarithmic scores of the form $-\log f(\boldsymbol{y}_{\mathcal{T}} \mid \boldsymbol{y}_{\mathcal{N}\setminus\mathcal{T}})$ could be calculated.

Again we can use the Monte Carlo estimated logarithmic scores to assess the Monte Carlo error of the exact sampling results. We have compared the analytical with the exact sampling log-scores for all examined models in the case studies. We found that the maximum absolute differences for the three sections were 0.017, 0.026 and 0.03, respectively, while the mean absolute differences were 0.002, 0.002 and 0.002. These Monte Carlo errors are very small compared to the approximate sampling errors, with maximum deviances 1.987, 1.007 and 1.357, and mean deviances 0.062, 0.062 and 0.054, respectively in the three sections when all models are pooled. For illustration we show comparison plots for three selected models in Figure 3.3. In all three panels, no clear deviations of points from the identity line can be reported. This is not surprising because the maximum and mean deviances are even lower here than for the one-step-ahead log-scores (cf. page 31), where already almost no large errors were visible in the comparison plots.

***Figure 3.3*** – *Comparison of analytical leave-one-out log-scores (x-axis) and corresponding exact sampling log-scores (y-axis), for three models from sections 3.4.2, 3.5.2 and 3.6.2.*



(a) Coal-mining data model 1 (b) Tokyo rainfall data model 2 (c) Nile discharge data model 3

### 3.3.3 Goodness-of-fit assessment

Obtaining samples $\boldsymbol{y}^*$ from the posterior-predictive distribution

$$f(\boldsymbol{y}^* \mid \boldsymbol{y}) = \iint f(\boldsymbol{y}^* \mid \boldsymbol{\xi}, \boldsymbol{\theta}) f(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\xi}\, d\boldsymbol{\theta}$$

is easy because $f(\boldsymbol{y}^* \mid \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{t\in\mathcal{N}} f(y_t^* \mid \boldsymbol{\xi}_t)$: In each iteration of the posterior sampling scheme from section 3.2.4 which produces a sample for $\boldsymbol{\xi}_t$, draw $y_t^*$ from the likelihood $f(y_t \mid \boldsymbol{\xi}_t)$, for all times $t \in \mathcal{N}$. The asterisk marks the random quantity $\boldsymbol{y}^*$ as a hypothetical *replicated* data vector, replicated from the original *observed* data vector $\boldsymbol{y}$: Generally this means that the replication $\boldsymbol{y}^*$ comes from the same model as the observed data $\boldsymbol{y}$, particularly that potential covariates $\boldsymbol{x}_t$ for the observation $y_t$ are identical for the replication $y_t^*$. However, the posterior-predictive distribution conditions on the observed data $\boldsymbol{y}$, and therefore issues a probabilistic forecast for a new independent replicate data set, based on the information in the observed data.

Afterwards, the estimated distribution $f(y_t^* \mid \boldsymbol{y})$ can be compared to the observed count $y_t$. Note that $y_t$ influences its predictive distribution directly. So these posterior-predictive checks are goodness-of-fit checks of our model rather than predictive checks. The corresponding question is how well our model can fit the *known* data, rather than how well our model can predict *new* data. The question could also be phrased "Is the model consistent with the data?", as Gelman, Carlin, Stern, and Rubin (2003, p. 159) call their posterior-predictive checking section.

If PIT histograms are produced from posterior-predictive samples, they usually show pictures typical for an overdispersed forecast. This is then a sign for a good fit of the probability model to the given data: if $y_t$ is fitted well by the model, the density $f(y_t^* \mid \boldsymbol{y})$ will be centered around the known $y_t$, thus producing a PIT value near 0.5. Very low or very high PIT values would point out badly fitted counts, which could be outliers in the data set. So for a well-fitting model we expect to see a hump-shaped histogram as in panel (c) on page 10.

Analogously, if we look at mean scores produced from posterior-predictive samples, they show us the goodness-of-fit of the considered model rather than the predictive performance for new data. In fact, there are even estimation approaches based on minimizing the mean score with respect to the model parameters (Gneiting and Raftery 2007, p. 374). Since we do not use the scores for estimating the model, we can use them for the goodness-of-fit assessment after the Bayesian model estimation, with smaller mean scores corresponding to a better fit of the model. Relatively large score values for single data points can point out badly fitted counts, analogously to extreme posterior-predictive PIT values.

### 3.3.4 Summary

The five different predictive sampling schemes are illustrated in Figure 3.4. The graphic emphasizes the difference between the two predictive assessments that we have introduced in this section: While the leave-one-out cross-validation assessment is a symmetric procedure which does not require the time series structure of the data, the one-step-ahead assessment is asymmetric and is only sensible because we know the ordering of the data points. The term "asymmetric" summarizes the fact that for each predicted observation, a different number of observed previous data points is utilized. If we reversed the time series, the result of the following one-step-ahead assessment could differ from those obtained from the original time series, except for the mean log-score and the marginal likelihood. Yet, we would get the same leave-one-out assessment result as before, also for the (continuous) ranked probability score. The approximate versions are distinguished from the corresponding exact assessment strategies by the fact that data points are now partially observed which were treated as unobserved by the exact versions. Finally, the posterior-predictive sampling differs from the leave-one-out sampling schemes in the predicted (replicate) observation being fully observed.

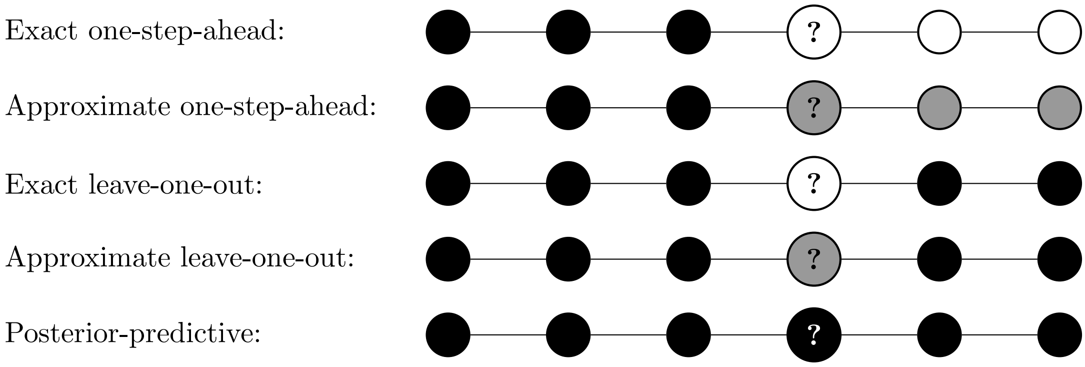


***Figure 3.4*** – *Summary graphic of the five sampling schemes, for an example of $n = 6$ nodes where the observation at $t = 4$ is predicted, which is symbolized by the question marks (?). The circles represent observed data (●), partially observed data (◍) and unobserved data (○). While observed data is used for both the change points sampling and the parameter levels sampling, partially observed data is only used for the change points sampling. Unobserved data is not used for the posterior sampling.*

We were able to assess the Monte Carlo error inherent to the logarithmic score estimation based on samples (cf. page 14). We have observed that using a sample size of 10 000 is well sufficient to very accurately calculate both one-step-ahead and leave-one-out logarithmic scores by exact sampling, without needing the marginal likelihood formula (3.2.10). The results are encouraging, because although we do have the marginal likelihood formula for the conjugate change point model, in more complex models it is usually not available, particularly if MCMC methods need to be employed to produce posterior samples. For example, Chapter 4 examines Bayesian normal random effects models which are fitted by Gibbs sampling. We can hope that the small Monte Carlo errors translate to that model family. Yet, also for the predictive assessment of the conjugate change point models the results are encouraging, because the (continuous) ranked probability scores really need to be estimated using samples, and now we can be more confident about their precision.

## 3.4 Poisson-Gamma model

The Poisson-Gamma change point model is described in section 3.4.1, which is a special case of the general framework from section 3.2. The proposed methodology from section 3.3 is applied in a case study using a data set on coal mining previously analyzed in the literature in section 3.4.2.

### 3.4.1 The special change point model

#### Data

The data form to which the Poisson-Gamma change point model may be applied is a time series $\boldsymbol{y}$ of counts $y_t \in \mathbb{N}_0$. For example, $y_t$ could be the infectious disease count in year $t$ in a certain region. In parallel, positive offsets $e_1, e_2, \dots, e_n$ are recorded. So the covariates are here $\boldsymbol{x}_t = e_t$. For the infectious disease count $y_t$, the number of susceptible persons in year $t$ could be relevant and be chosen to be the offset $e_t$.

#### Model

We assume independent Poisson distributions for the counts $y_t$ with rates $\lambda_t$ relative to the offsets $e_t$:

$$y_t \,|\, \lambda_t, e_t \overset{ind}{\sim} \mathrm{Po}(e_t \lambda_t), \quad t \in \mathcal{N}.$$

So the parameters are scalar for this model ($\boldsymbol{\xi}_t = \lambda_t$) and the response density is $f(y_t \,|\, \boldsymbol{\xi}_t, \boldsymbol{x}_t) = \mathrm{Po}(y_t \,|\, e_t \lambda_t)$.

#### Prior

As described on page 21, we need to select a conjugate parameters prior. Since we have specified a Poisson likelihood for the observed counts, the gamma distribution must be the prior for the rate parameters. So for the $k+1$ rate levels independent identical gamma priors with hyperparameters $\alpha, \beta > 0$ are specified,

$$\lambda^{(j)} \overset{iid}{\sim} \mathrm{G}(\alpha, \beta), \quad j = 1, \dots, k+1.$$

The hyperparameter $\boldsymbol{\phi}$ has elements $\alpha, \beta$ here, and $f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{\phi}) = \mathrm{G}(\lambda^{(j)} \,|\, \alpha, \beta)$.

For the block marginal likelihood (3.2.6) we have

$$\begin{aligned}
f_{block}(\boldsymbol{y}_{\mathcal{S}}) &= \int\limits_{\mathbb{R}_+} \prod_{t\in\mathcal{S}} \mathrm{Po}(y_t \,|\, e_t\lambda^{(j)}) \cdot \mathrm{G}(\lambda^{(j)} \,|\, \alpha, \beta)\, d\lambda^{(j)} \\
&= \int\limits_{\mathbb{R}_+} \prod_{t\in\mathcal{S}} \frac{(e_t\lambda^{(j)})^{y_t}}{y_t!} \exp(-e_t\lambda^{(j)}) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} (\lambda^{(j)})^{\alpha-1} \exp(-\lambda^{(j)}\beta)\, d\lambda^{(j)} \\
&= \prod_{t\in\mathcal{S}} \frac{e_t^{y_t}}{y_t!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \int\limits_{\mathbb{R}_+} (\lambda^{(j)})^{\sum_{t\in\mathcal{S}} y_t+\alpha-1} \exp\left(-\lambda^{(j)} \left[\sum_{t\in\mathcal{S}} e_t + \beta\right]\right) d\lambda^{(j)} \\
&= \prod_{t\in\mathcal{S}} \frac{e_t^{y_t}}{y_t!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\sum_{t\in\mathcal{S}} y_t + \alpha)}{(\sum_{t\in\mathcal{S}} e_t + \beta)^{(\sum_{t\in\mathcal{S}} y_t+\alpha)}}.
\end{aligned}$$

While this is not exactly a Poisson-Gamma density (because different counts $y_t$ share the same rate $\lambda^{(j)}$), the derivation of this block density is analogue to the derivation of a Poisson-Gamma density.

### Posterior

In order to sample the rate parameters given the change points, we need the block posterior density (3.2.7). Its form can be derived from the product of the block Poisson likelihood and the Gamma prior for the rate level:

$$\begin{aligned}
f_{block}(\lambda^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}, \alpha, \beta) &\propto \prod_{t\in\mathcal{S}} \mathrm{Po}(y_t \,|\, e_t\lambda^{(j)}) \cdot \mathrm{G}(\lambda^{(j)} \,|\, \alpha, \beta) \\
&\propto (\lambda^{(j)})^{\sum_{t\in\mathcal{S}} y_t+\alpha-1} \exp\left(-\lambda^{(j)} \left[\sum_{t\in\mathcal{S}} e_t + \beta\right]\right) \\
&\propto \mathrm{G}\left(\lambda^{(j)} \,|\, \sum_{t\in\mathcal{S}} y_t + \alpha, \sum_{t\in\mathcal{S}} e_t + \beta\right).
\end{aligned}$$

## 3.4.2 Case study

We use the data set on coal-mining disasters in Great Britain which has been introduced by Maguire, Pearson, and Wynn (1952) and has been extended the last time by Raftery and Akman (1986). In its present form, the data set gives the time interval in days between explosions in British coal mines involving 10 or more fatalities, from the beginning of 1851 until the end of 1962. This data has been used frequently in the literature, e. g. by Denison, Holmes, Mallick, and Smith (2002, p. 179). We will examine the number of disasters per year; the total number is 191. The time series of length $n = 112$ is plotted in Figure 3.5.

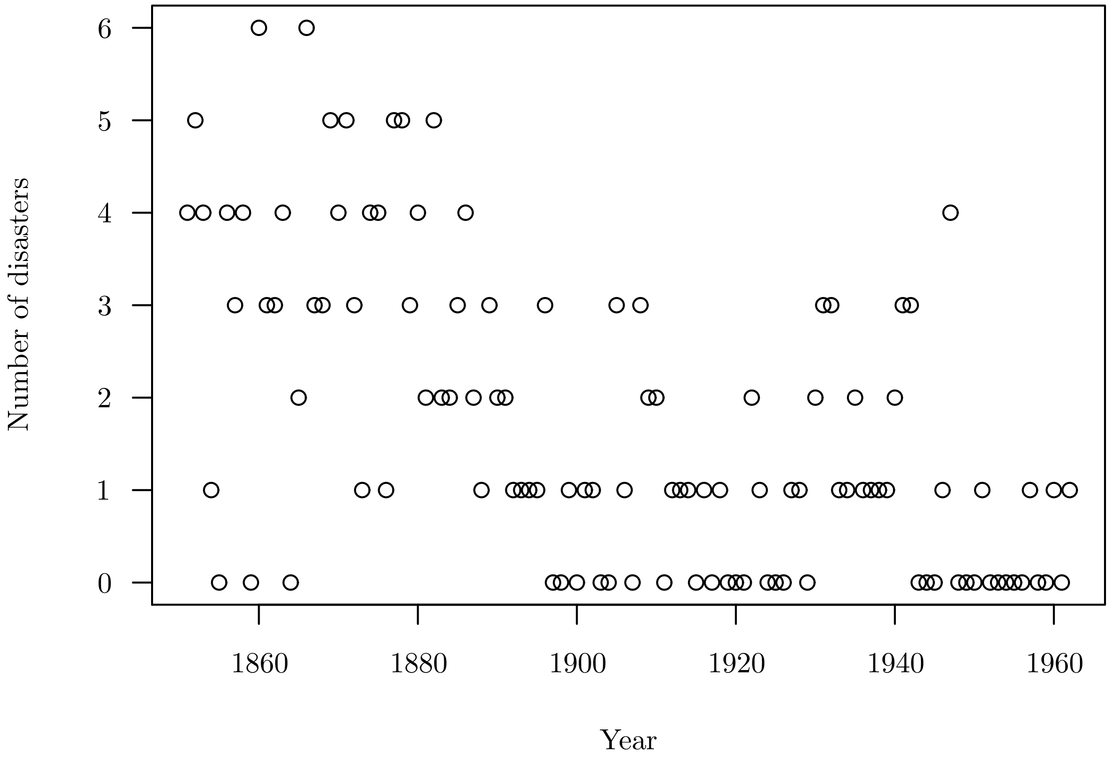


***Figure 3.5*** – *Coal-mining disasters data: Number of disasters (with at least 10 fatalities) per year.*

### Model fitting

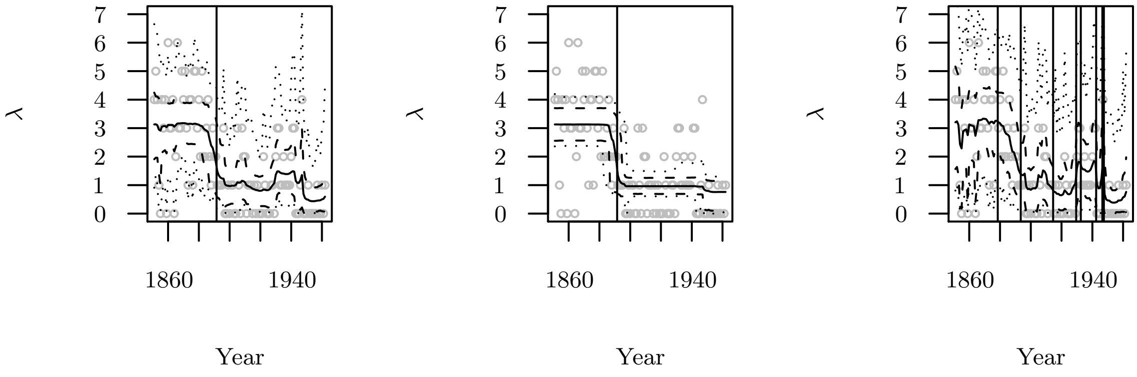
The assumption of Poisson distributions for the yearly disaster counts $y_t$ is sensible, because we have aggregated events in fixed length time intervals. Unfortunately we do not have the time series of the number of working coal mines in Great Britain, so we can not use them as offsets. Instead we set $e_t \equiv 1$ for all times $t$. This means that e. g. higher rates could possibly be due to more working coal mines with comparable risks of explosions, and need not be evidence for higher risks in the same coal mines.

The first model we will fit to the data uses the flat number prior for the change points, and hyperparameters $\alpha = 1.7054, \beta = 1$ for the rates prior, such that the prior mean and variance equal the average disasters count 1.7054. This model was also considered by Hofmann (2007, p. 26).

The second model we want to assess also uses the flat number prior for the change points, but with hyperparameters $\alpha = 0.017054, \beta = 0.01$ for the rates prior. So the prior mean of the rates still equals 1.7054, but the variance is now 170.54, leading to a vaguer prior.

The last model we consider uses the binomial number prior with probability $\pi = 0.2$ for a change point between any two years of the time series. The rates prior hyperparameters are chosen as for the first model ($\alpha = 1.7054, \beta = 1$).

***Figure 3.6*** – *Posterior rates trends for the three change point models. Pointwise HPD (dashed lines) as well as simultaneous (dotted lines) 95% credible intervals, which were estimated by simulating 10 000 samples, for the rates trend are given. The change point locations in the respective MAP models are marked with vertical lines.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

We have produced 10 000 samples each from the posterior distributions. The estimated rates trends and the change point locations in the MAP model are shown in Figure 3.6. The simultaneous credible bands for the rates trends were computed with the quantile method of Besag, Green, Higdon, and Mengersen (1995, p. 30).

The differences between the model fits are interesting: Both model 1 in panel (a) and model 2 in panel (b) have their MAP model change points at $t = 41$, which corresponds to the year 1891. The posterior probabilities for these configurations are $5.28 \cdot 10^{-3}$ and $1.72 \cdot 10^{-1}$, respectively. Yet, the posterior rates trend averaged over the change point configurations is much more variable for model 1 than for model 2. This is in fact an example where a vaguer hyperprior leads to a more parsimonious model, a phenomenon known as Lindley's paradox (Lindley 1957). On the other hand, model 3 in panel (c) exhibits an even more wiggly rates trend, and its MAP model with probability $1.29 \cdot 10^{-11}$ contains 8 change points after the years 1878, 1893, 1914, 1929, 1932, 1942, 1946 and 1947. This model obviously overfits the data.

Note that the posterior probability mass of the respective change points distribution $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ is much more spread out to different change points configurations for model 3 than for model 1, and also more for model 1 than for model 2: this can be seen from the MAP model probabilities of the best configurations found in the respective samples $\boldsymbol{\theta}_{[1]}, \ldots, \boldsymbol{\theta}_{[10\,000]}$. A consequence is that the exploration of the posterior should be easiest for model 2, and hardest for model 3. For the latter, model averaging is extremely important in order not to trust a questionable best change points configuration – running the sampler again could lead to a totally different MAP model configuration.

The log marginal likelihood values $\log f(\boldsymbol{y})$ of the three change point models are $-177.487$, $-187.21$ and $-177.986$, respectively. So if we should decide on the basis of the log marginal likelihood, model 1 would be our best choice. Yet, we want to examine the calibration and predictive capabilities of the three models before making a final decision.

### One-step-ahead predictive assessment

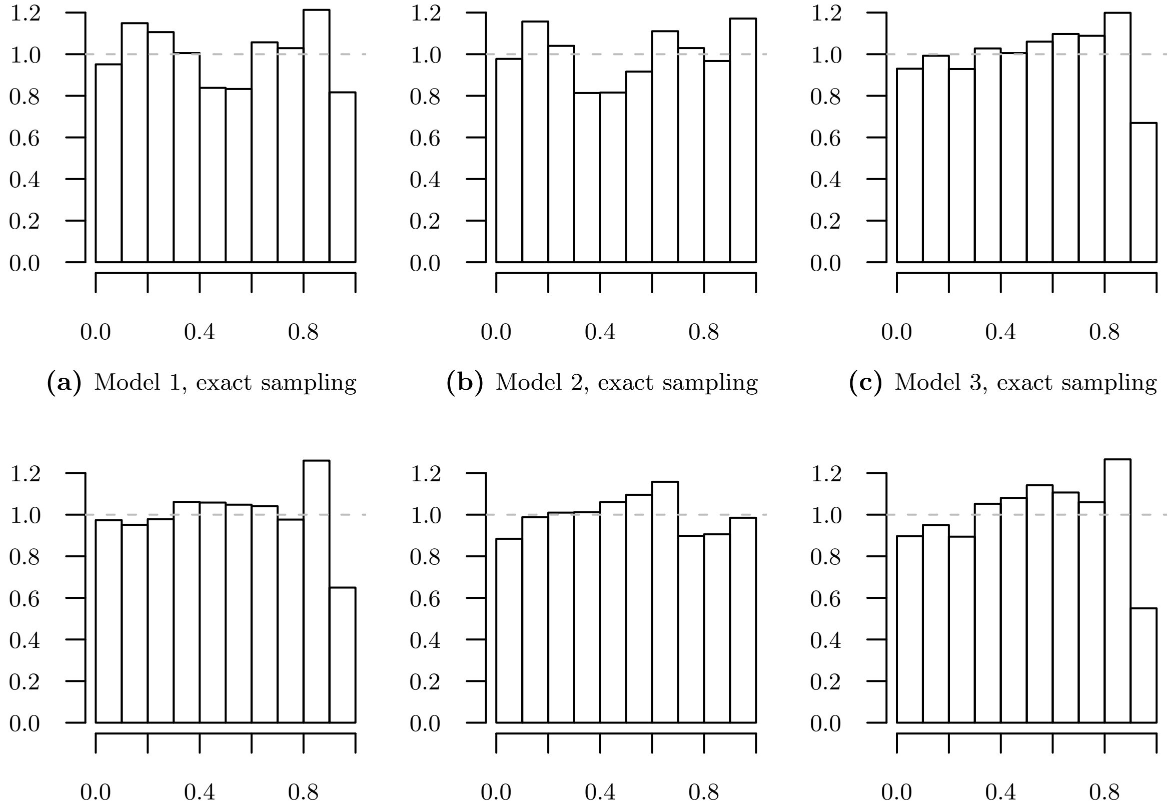
For practical purposes, good one-step-ahead prediction is especially important. We want to check that for the three models in question using the sampling tools from section 3.3.1, and the PIT and scores from chapter 2.

First, we generate 10 000 rates samples, both from the exact and the approximate one-step-ahead predictive distributions, for all three models. That is, for each model, and for all last times $t = 0, 1, \dots, n-1 = 111$, we sample 10 000 variates exactly from $f(\lambda_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$ and again 10 000 variates from the approximation $\tilde{f}(\lambda_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$. For the one-step-ahead sampling of the next rate given a change point configuration, we use the sequential approach: First the change point occurrence before the next time is sampled, and then the rate is either drawn from the prior or set to the last rate in the observed time series. Altogether, this takes 119, 62, 155 seconds for the exact sampling and 75, 16, 107 seconds for the approximate sampling, for the three different models, respectively.

Second, we plug each rate sample $\lambda_t$ into the Poisson likelihood and keep one Poisson variate $y_t^* \sim \mathrm{Po}(\lambda_t)$ as a sample from the (approximated) one-step-ahead predictive distribution $F_t$ for time $t$ given all prior times.

Then, we estimate the PIT values $F_t(y_t)$ and the "pre-PIT" values $F_t(y_t - 1)$, for $t = 1, 2, \dots, n$, using the empirical distributions $\hat{F}_t$ of the one-step-ahead predictive samples and the true observations $y_t$. This is done for both sampling approaches and for all three models, and results in the PIT histograms shown in Figure 3.7. Overall, all the models look well calibrated. Only for model 3 in panel (c), a tendency towards too few PIT probability in the last bin $[0.9, 1.0]$ is observed. So the fraction of true observations falling into the upper 10% prediction intervals is smaller than 10%, which is the fraction we expected for perfectly calibrated predictive distributions. Therefore, the upper 10% prediction intervals are too large. This is an argument for a slight overdispersion of model 3. This tendency is still visible in panel (f) under the approximate sampling scheme, as the whole histogram looks very similar to its exact counterpart. For model 1 and model 2 the histograms also match quite well, even if the approximate panel (e) speaks for a bit better calibration of model 1 than the exact panel (b).

***Figure 3.7*** – *PIT histograms for calibration assessment of the one-step-ahead prediction in the three change point models (columns). The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*



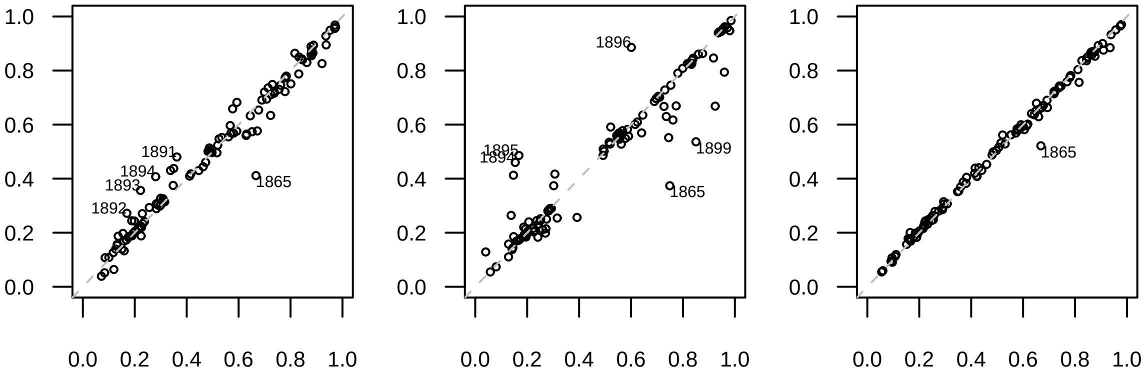
**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

The estimated mid-PIT values $0.5\big(F_t(y_t - 1) + F_t(y_t)\big)$, $t = 1, 2, \ldots, n$, which were introduced in section 2.2.1 are compared between the exact and the approximate sampling schemes in Figure 3.8. For model 1, most departures of the approximate PIT values from their exact counterparts occur around the probable change point 1891 in this model, see panel (a). For model 2 in panel (b), there are more and greater differences: one might suspect that some are related to the second step around 1950 which is visible in the model-averaged fit in panel (b) of Figure 3.6. It is interesting that for the overfitting model 3 in panel (c), only for a single year a relevant deviation of the approximation is observed.

Now we turn to proper scoring rules. We estimated the ranked probability scores $RPS(F_t, y_t)$ for $t = 1, 2, \ldots, n$. Moreover, we estimated the logarithmic scores $LogS(F_t, y_t)$. For the prediction time $t$, we used the rate samples $\lambda_{t,[j]}, j = 1, \ldots, m = 10\,000$ for the

***Figure 3.8*** – *Comparison of exact (x-axis) and approximate (y-axis) mid-PIT values for calibration assessment of the one-step-ahead prediction in the three change point models. At most 5 time points where the absolute difference between the two values exceeds 0.1 are labelled.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

Monte Carlo estimate

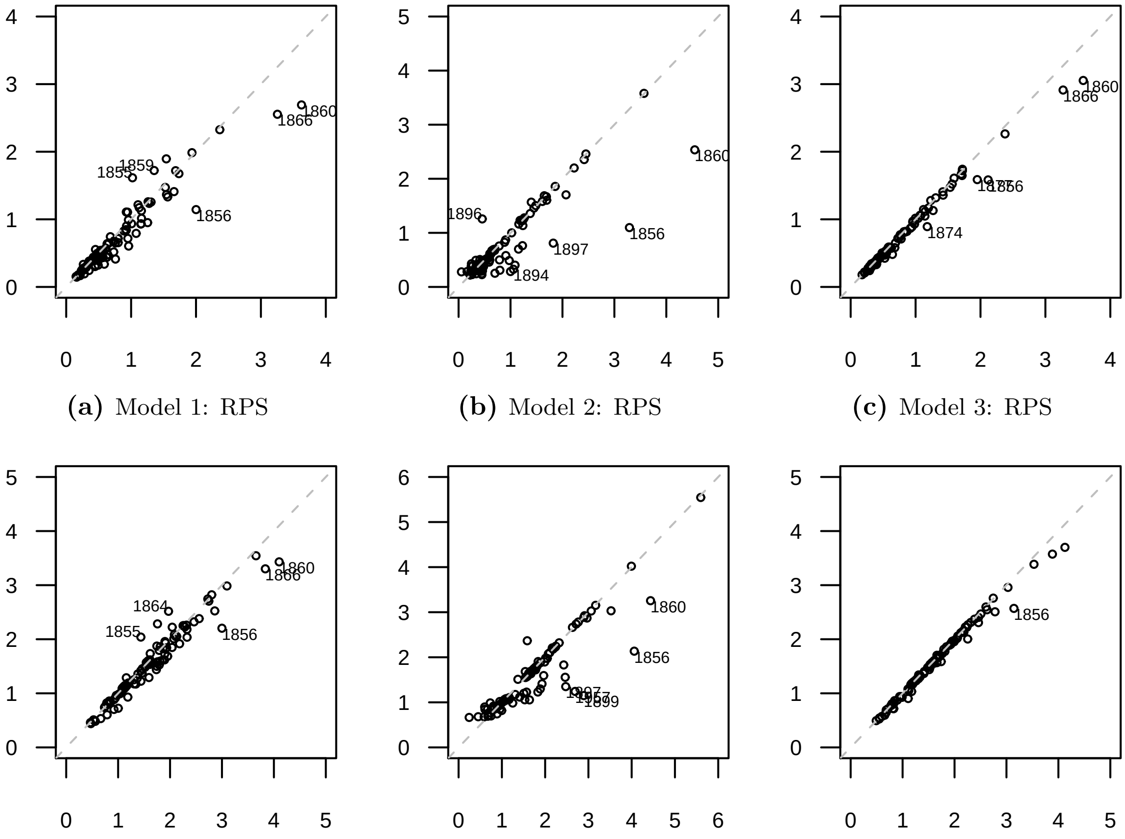
$$\widehat{LogS}(F_t, y_t) = -\log\left\{\frac{1}{m}\sum_{j=1}^{m}\text{Po}(y_t \mid \lambda_{t,[j]})\right\}.$$

The exact and approximate scores of both proper scoring rules are compared in Figure 3.9. For model 1, only for a few years at the beginning of the time series there are larger discrepancies. For the log-scores in panel (d) the overall picture is similar to the RPS in panel (a). The same can be said about model 3 in panels (c) and (f), while the absolute deviations of the approximate score values are even smaller. Yet, for model 2 in panels (b) and (e), there are more larger deviations than in model 1, and the approximations do not seem to work very well.

The differences of the approximate and exact mid-PIT values, ranked probability and logarithmic scores are plotted against the time in Figure 3.10. Here we see more clearly where in time large approximation errors occur: especially at the beginning of the time series, when the difference between the exact and the approximate scheme is largest, and around the big step in the rates trend before the turn of the century. It is difficult to approximate the values from model 2, and easier for model 3. We expect the differences to get smaller at the end of the time series, because then the exact and the approximate sampling scheme share more common data. The RPS differences meet our expectations, but there are some big mid-PIT and log-score differences near the last times.

The mean scores for the proper scoring rules assessment of the one-step-ahead prediction are summarized in Table 3.1. Looking at both RPS rows in the table, it is not

***Figure 3.9** – Comparison of exact (x-axis) and approximate (y-axis) scores for one-step-ahead prediction in the three change point models (columns). The panels in the upper row compare the RPS values, while the panels in the lower row compare the log-scores. Time points with the 5 largest absolute differences between the exact and approximate score values exceeding 0.25 (RPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: RPS

**(b)** Model 2: RPS

**(c)** Model 3: RPS

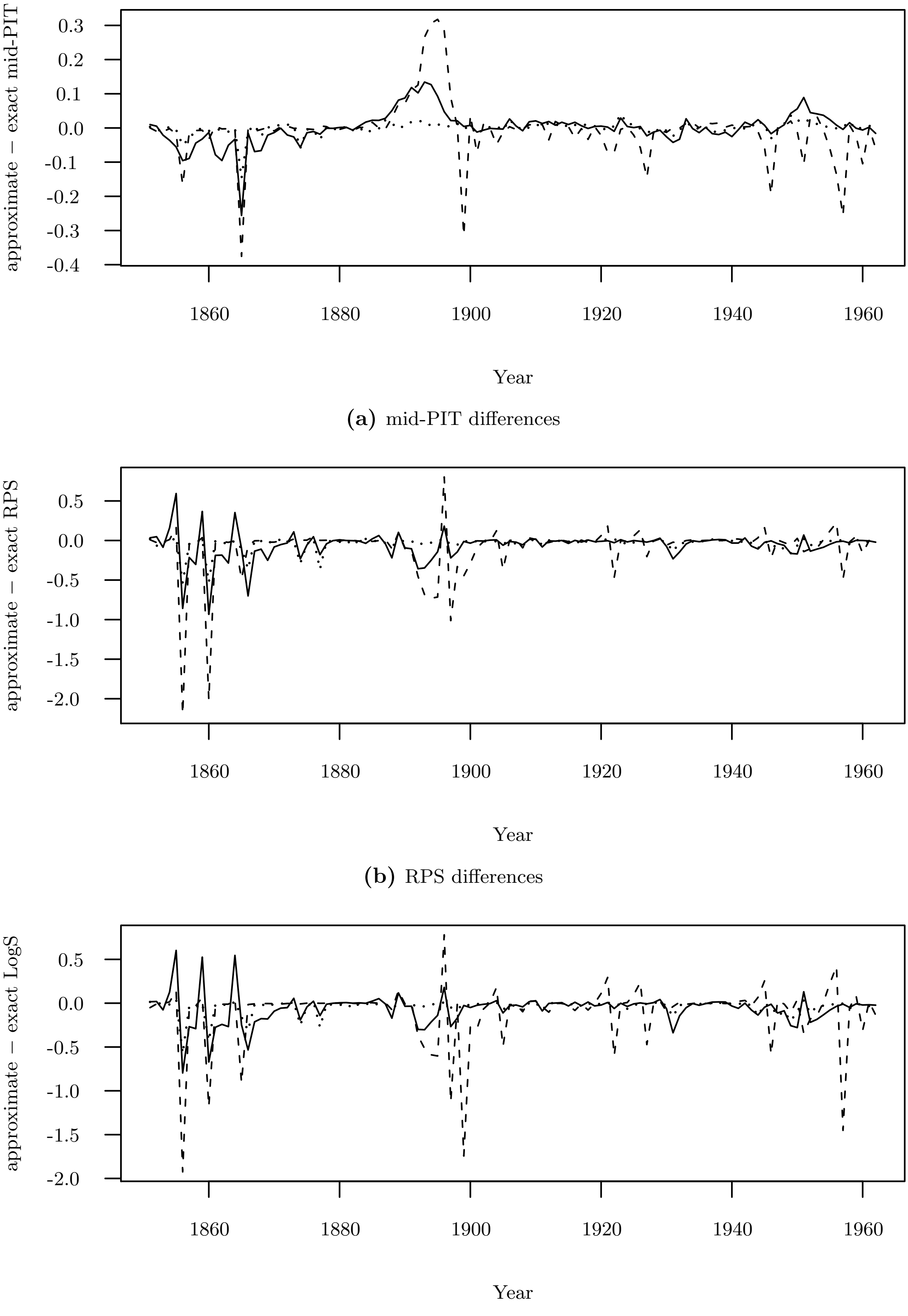
**(d)** Model 1: log-score

**(e)** Model 2: log-score

**(f)** Model 3: log-score

surprising that the paired permutation test comparing the RPS values from the exact and approximate approaches gives estimated p-values $2 \cdot 10^{-4}$ for model 1, $9 \cdot 10^{-4}$ for model 2 and $1 \cdot 10^{-4}$ for model 3. That is, in almost all of the 10 000 sampled permutations of the value pairs, the resulting mean score differences were smaller than the observed differences. This suggests that the approximate assessment is conservative and underestimates the RPS which is a generalized prediction error. If we had the exact sampling results for the RPS, we would choose model 1 or model 3. If we had the approximate sampling results, we would prefer model 1. If we directly compare the exact and approximate log-scores of each model, we get p-values $4 \cdot 10^{-4}$ for model 1, $1.8 \cdot 10^{-3}$ for model 2 and $1 \cdot 10^{-4}$ for model 3 of the paired permutation test. So also for this proper scoring rule, the approximate sampling significantly underestimates the exact sampling mean scores.

***Figure 3.10*** – *Differences of the approximate and exact mid-PIT values, ranked probability and logarithmic scores for the one-step-ahead prediction, for model 1 (——), model 2 (- - -) and model 3 (........).*



**(a)** mid-PIT differences

**(b)** RPS differences

**(c)** Log-score differences

For this example if we use the logarithmic score, both the exact and the approximate method assign model 1 the best scores – the ranking is the same under the approximate method.

***Table 3.1*** – *Mean ranked probability and logarithmic scores for the one-step-ahead prediction of the three models, under the exact and approximate sampling schemes.*

| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| RPS | exact | 0.74 | 0.80 | 0.74 |
| | approximate | 0.68 | 0.71 | 0.70 |
| log-score | exact | 1.58 | 1.67 | 1.59 |
| | approximate | 1.52 | 1.56 | 1.55 |

### Leave-one-out predictive assessment

Next, we will do a cross-validation assessment where we leave one observation out in each iteration, which is the leave-one-out strategy. How good are the models at predicting the missing observation? And how close are the exact and approximate model assessment results?

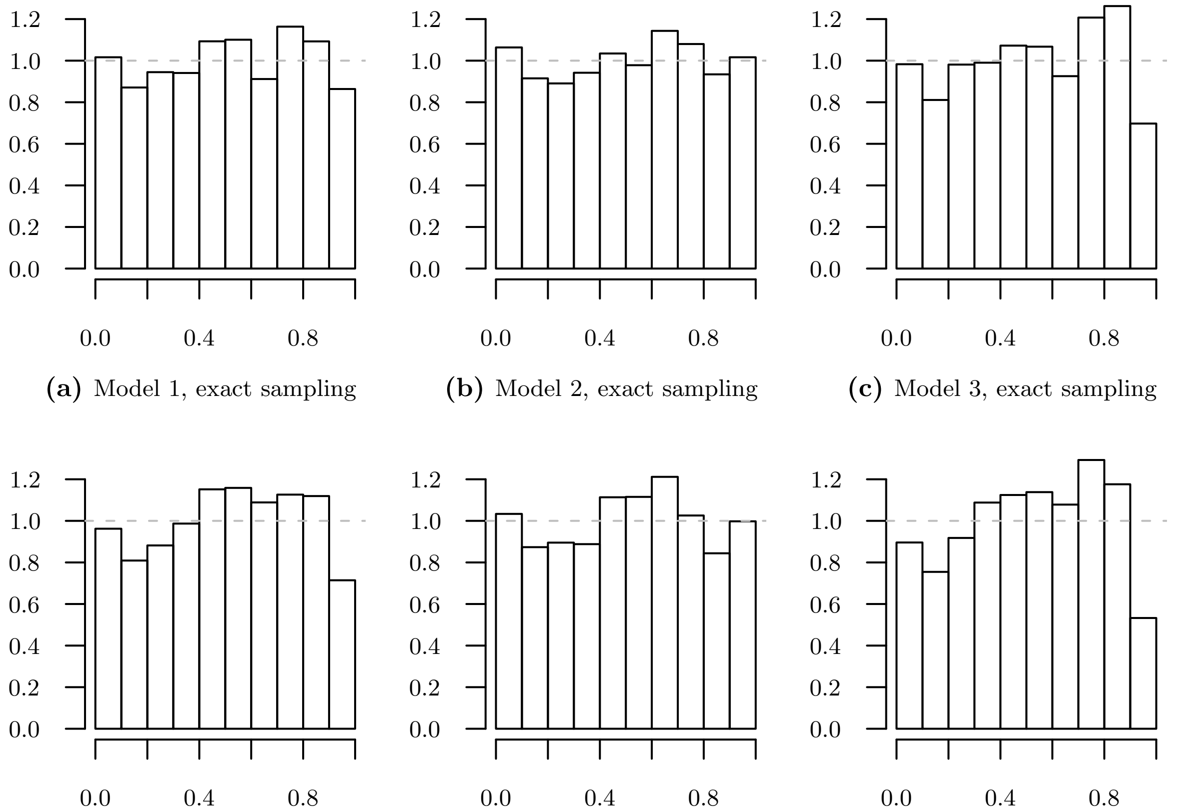
First, we generate 10 000 rates samples, both from the exact and the approximate leave-one-out distributions, for all three models. That is for each model, and for each time $t = 1, 2, \ldots, n = 112$, we sample 10 000 variates exactly from $f(\lambda_t \mid \boldsymbol{y}_{N\setminus\{t\}})$ and again 10 000 variates from the approximation $\tilde{f}(\lambda_t \mid \boldsymbol{y}_{N\setminus\{t\}})$. Altogether, this takes 258, 146, 269 seconds for the exact sampling and 124, 23, 131 seconds for the approximate sampling, for the three different models, respectively.

Second, we plug each rate sample $\lambda_t$ into the Poisson likelihood and keep one Poisson variate $y_t^* \sim \text{Po}(\lambda_t)$ as a sample from the (approximated) leave-one-out predictive distribution $F_t$ for time $t$ given all other times.

The PIT histograms are shown in Figure 3.11. Model 1 and model 2 look well calibrated if we judge them by panel (a) and panel (b), respectively. The approximate results in panels (d) and (e) are similar to their exact counterparts. For model 3 in panel (c) we see again a tendency towards overdispersion, which is even more pronounced in the approximate panel (f).

The mid-PIT values $0.5\big(F_t(y_t - 1) + F_t(y_t)\big)$, $t = 1, \ldots, n$, are compared between the exact and approximate sampling schemes in Figure 3.12. For model 1 and model 3, no large deviation of an approximate PIT value from the exact PIT value is noticeable in panels (a) and (c), respectively. For model 2 in panel (b), only two cross-validation PIT

***Figure 3.11*** – *PIT histograms for calibration assessment of the leave-one-out prediction in the three change point models (columns). The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*



**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

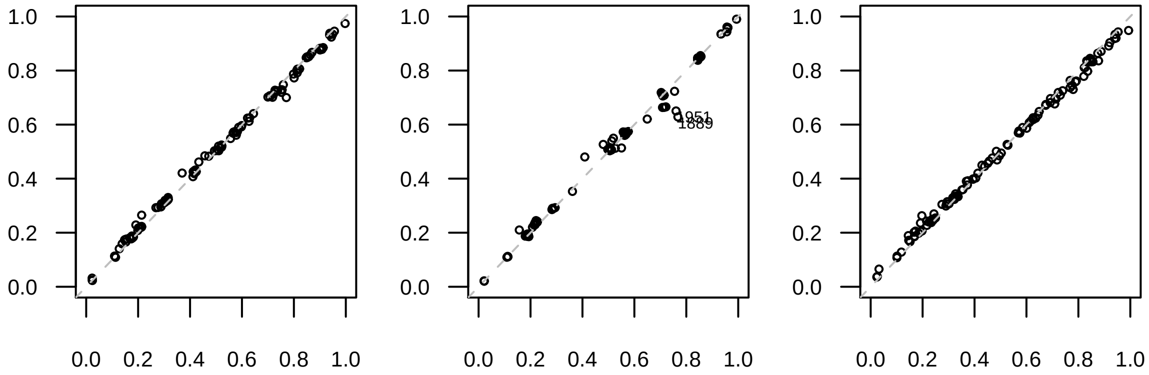
**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

values can not be well approximated: these are again around the two steps from panel (b) in Figure 3.6.

The exact and approximate scores of both proper scoring rules are compared in Figure 3.13. The RPS plots in the upper row look very similar to the log-score plots in the lower row. Altogether, the exact RPS values are well approximated, except for the year 1947, which is labelled in all but one plot.

The mean scores for the proper scoring rules assessment of the leave-one-out prediction are summarized in Table 3.2 on page 49. For the RPS values, the paired permutation test still clearly rejects the hypotheses of same location parameters for approximate and exact scores. However, the absolute differences between approximate and exact means are smaller than for the one-step-ahead predictive assessment. If we had the exact sampling results, we would choose model 1, and if we had the approximate sampling results, we

***Figure 3.12*** – *Comparison of exact (x-axis) and approximate (y-axis) mid-PIT values for calibration assessment of the leave-one-out prediction in the three change point models. At most 5 time points where the absolute difference between the two values exceeds 0.1 are labelled.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

would choose model 3. However, the differences are quite small, and the mean model 1 scores are close to the mean model 3 scores under both sampling schemes. The differences between exact and approximate mean log-scores are significant in the paired permutation test for all models. If we look at the logarithmic score ranking of the models, we are again undecided about whether model 1 or model 3 has the best predictive performance. This result is the same under exact and approximate assessment.

### Posterior-predictive checking

For comparison with the one-step-ahead and leave-one-out predictive assessments, we will look at the results of posterior-predictive model checking.

As was described in section 3.3.3, we just plug in the posterior rate samples $\lambda_t$ into the Poisson likelihood to obtain samples $y_t^* \sim \text{Po}(\lambda_t)$ from the posterior-predictive distribution $F_t$, for times $t = 1, 2, \ldots, n$. So a big advantage of these checks is that they do not require the costly one-step-ahead- or leave-one-out-sampling of the rates, only sampling from the likelihood is necessary in addition to the model fitting.

The PIT histograms are shown in Figure 3.14. Only model 2 in panel (b) shows a “good” PIT histogram, while model 1 in panel (a) and especially model 3 in panel (c) show overdispersion. This result has been expected from the rates trends in Figure 3.6: While model 2 has a smooth fit to the given data, its rates trend does not follow every extreme observed count. Model 1 follows the given data more closely, and model 3 already overfits the given data. Therefore we have expected that model 3 fits the given data “best” in the sense of “closest”, so that its posterior-predictive PIT histogram will have the most hump-shaped form of all three models.

The mid-PIT values from the exact leave-one-out and the posterior-predictive sampling schemes are compared in Figure 3.15. The panels look similar to those in Figure 3.12, but there is a curvature in the point clouds which reveals the conservativeness of the posterior-predictive PIT values: those times where the exact PIT values are smaller than 0.5, the posterior-predictive PIT values are too large; and vice versa they are too small, where the exact PIT values are greater than 0.5. This shrinkage is strongest for the overfitting model 3, and weakest for the rather underfitting model 2. Note that the curvature is much less visible for the approximate PIT values in Figure 3.12.

The exact leave-one-out scores are compared with the posterior-predictive scores in Figure 3.16. For model 1 and model 2 in panels (a), (d) and (b), (e), the differences to the respective plots in Figure 3.13 are not very large. This bias of the posterior-predictive scores is more pronounced for model 3. This observation can also be explained by the fact that model 3 fits the given data most closely among all three models, so that its posterior-predictive distributions differ more from the corresponding exact leave-one-out distributions than for the other models.

The mean scores are summarized and compared to the leave-one-out scores in Table 3.2. Using the posterior-predictive model scores, model 3, the model with the most variable fit, appears to have the best fit to the data. Note that the absolute values are much smaller than in the leave-one-out assessment. Also based on the log-scores, model 3 scores best among the three models, in the posterior-predictive check. The absolute values are lower than for the exact and also for the approximate leave-one-out assessment. Following the concept from section 3.3.3, this means that the goodness-of-fit is best for model 3, and worst for model 2. Again, this ranking could have been expected from Figure 3.6.

***Table 3.2*** – *Mean ranked probability and logarithmic scores for the three models, under the exact and approximate leave-one-out and the posterior predictive sampling schemes.*

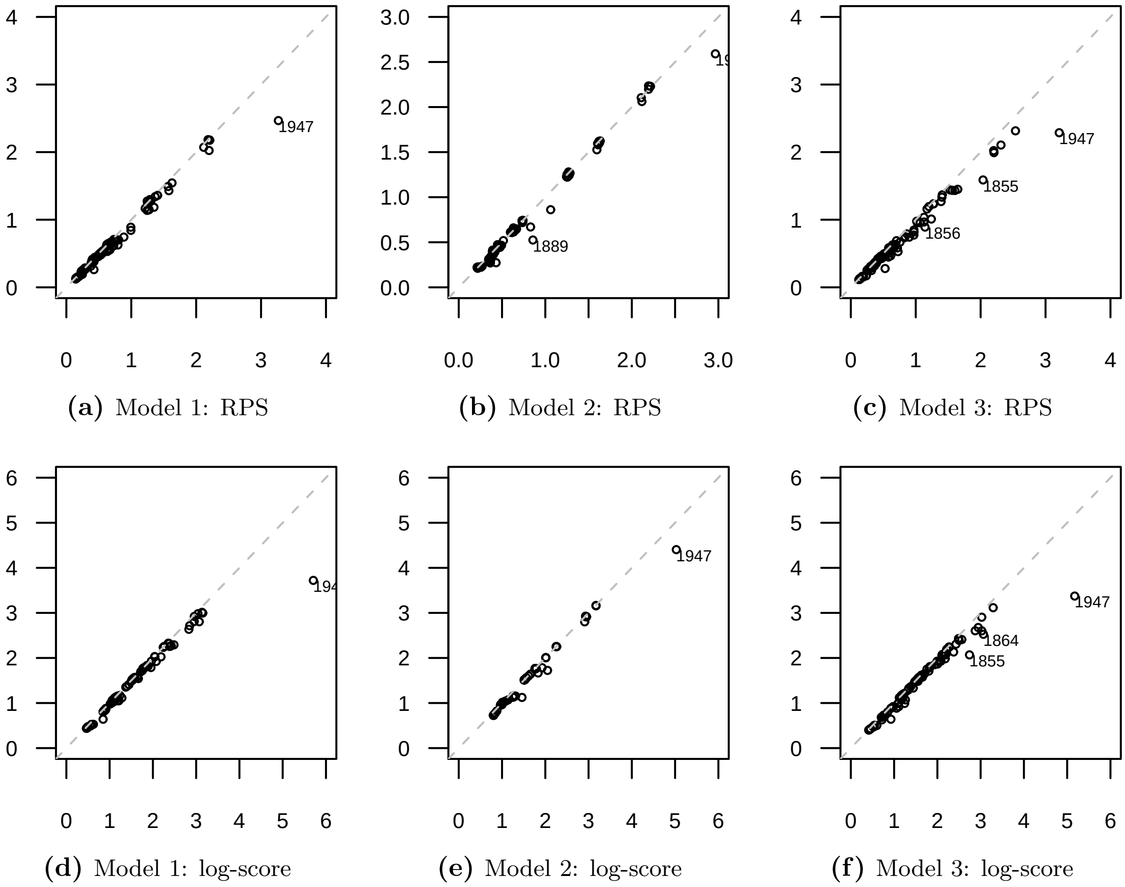
| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| RPS | exact leave-one-out | 0.67 | 0.68 | 0.68 |
| | approximate leave-one-out | 0.63 | 0.66 | 0.62 |
| | posterior-predictive | 0.59 | 0.64 | 0.54 |
| log-score | exact leave-one-out | 1.53 | 1.54 | 1.53 |
| | approximate leave-one-out | 1.46 | 1.51 | 1.45 |
| | posterior-predictive | 1.41 | 1.49 | 1.35 |

### Results

Overall, model 1 and model 3 are close in their predictive performances and log-marginal likelihoods. Model 3 can fit the data more closely, and is thus favored by the posterior-predictive checks, which means that its goodness-of-fit is best among all three considered models. Model 1 has better calibration and wins if the log-scores are used. Model 2 has a more parsimonious fit, but misses the rise in disasters between 1930 and 1940, and so shows worse mean scores in all five applied sampling schemes.

The approximate predictive assessment showed very similar results compared to the exact assessment, with only slight tendency to favoring the most complex model 3. In particular we have seen that the approximate results are much closer to the exact results than the posterior-predictive results. So the conservativeness of the posterior-predictive checks could be remedied to large extent by applying the approximate sampling scheme, while still saving much computational effort.

***Figure 3.13*** – *Comparison of exact (x-axis) and approximate (y-axis) scores for leave-one-out prediction in the three change point models (columns). The panels in the upper row compare the RPS values, while the panels in the lower row compare the log-scores. Time points where the absolute difference between the exact and approximate score values exceeds 0.25 (RPS) or 0.5 (log-score) are labelled (at most 5 points in each panel).*



**(a)** Model 1: RPS

**(b)** Model 2: RPS

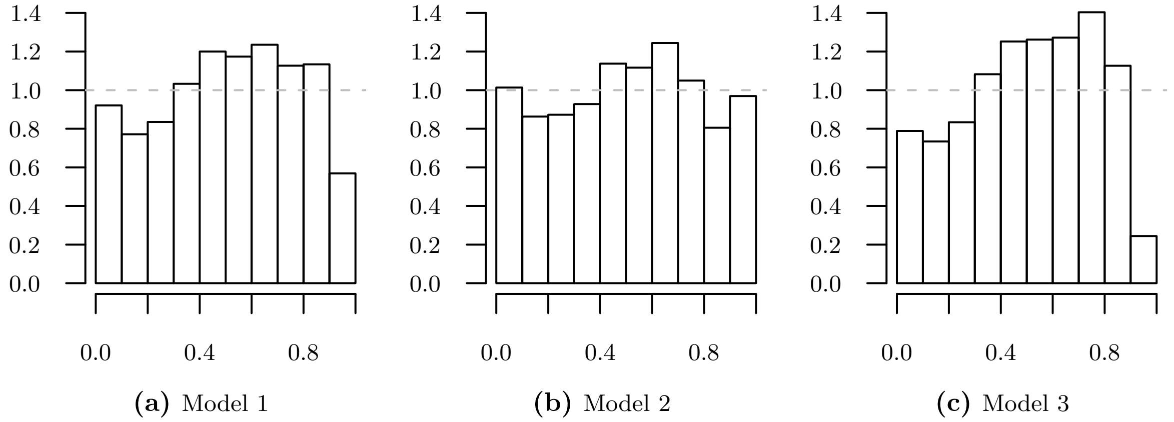
**(c)** Model 3: RPS

**(d)** Model 1: log-score

**(e)** Model 2: log-score

**(f)** Model 3: log-score

***Figure 3.14*** *– PIT histograms for posterior-predictive checking of the three change point models.*

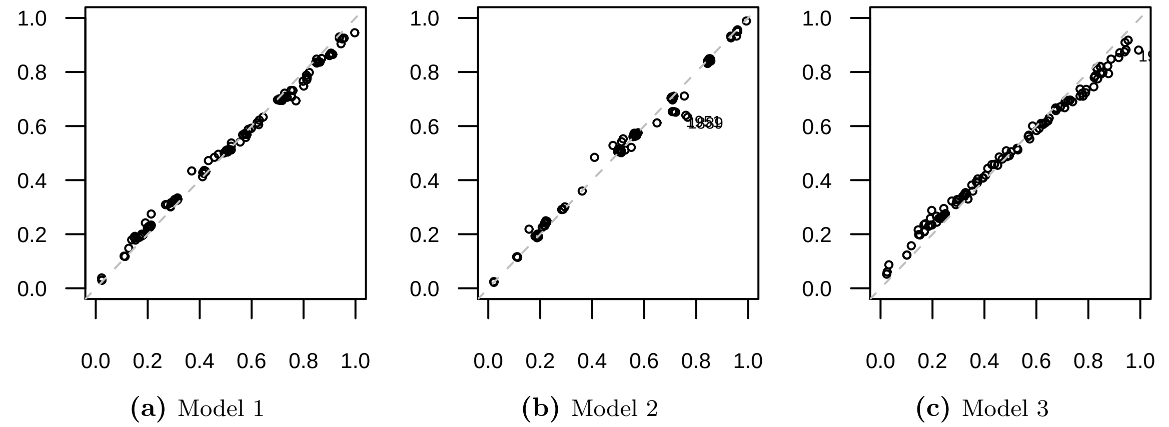


**(a)** Model 1

**(b)** Model 2

**(c)** Model 3

***Figure 3.15*** *– Comparison of exact leave-one-out (x-axis) and posterior-predictive (y-axis) mid-PIT values in the three change point models. Time points where the absolute difference between the two values exceeds 0.1 are labelled.*

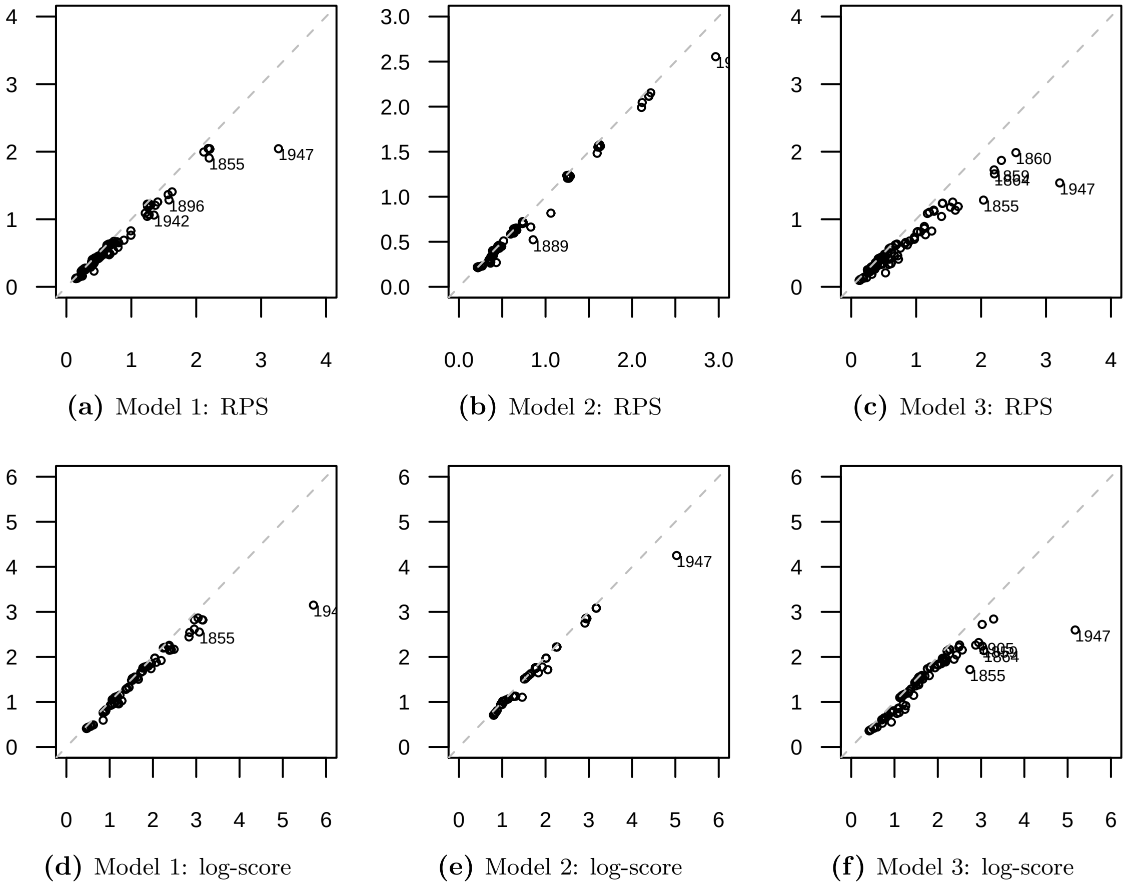


**(a)** Model 1

**(b)** Model 2

**(c)** Model 3

***Figure 3.16*** – *Comparison of exact leave-one-out (x-axis) and posterior-predictive (y-axis) scores in the three change point models (columns). The panels in the upper row compare the RPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact leave-one-out and posterior-predictive score values exceeds 0.25 (RPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: RPS

**(b)** Model 2: RPS

**(c)** Model 3: RPS

**(d)** Model 1: log-score

**(e)** Model 2: log-score

**(f)** Model 3: log-score

## 3.5 Binomial-Beta model

Another special case of the general conjugate change point model from section 3.2 is the combination of the binomial likelihood and the beta prior, which is presented in section 3.5.1. The case study in section 3.5.2 assesses three different Binomial-Beta change point models for the Tokyo rainfall data, using the five different predictive sampling schemes from section 3.3.

### 3.5.1 The special change point model

#### Data

The Binomial-Beta model is also suitable for count data $\boldsymbol{y} := (y_1, y_2, \ldots, y_n)$ like the Poisson-Gamma model, but here the maximum counts $n_1, n_2, \ldots, n_t$ which could have been observed must be available. So $y_t \in \{0, 1, \ldots, n_t\}$ is observed. For example, $y_t$ could be the number of pupils passing the Abitur in year $t$ in a certain school. Then $n_t$ is the number of pupils writing the Abitur in year $t$ in this school. The covariates are here the maximum counts, or sample sizes, $\boldsymbol{x}_t = n_t$.

#### Model

We assume independent binomial distributions with probabilities $\pi_t$ for the independent $n_t$ trials:

$$y_t \,|\, \pi_t, n_t \overset{ind}{\sim} \operatorname{Bin}(n_t, \pi_t), \quad t \in \mathcal{N}.$$

The model parameters are thus scalar for this model ($\boldsymbol{\xi}_t = \pi_t$) and the response density is $f(y_t \,|\, \boldsymbol{\xi}_t, \boldsymbol{x}_t) = \operatorname{Bin}(y_t \,|\, n_t, \pi_t)$.

#### Prior

The beta distribution is conjugate to the binomial likelihood, which we have chosen. So in order to get a conjugate change point model, we specify independent identical beta priors with hyperparameters $\alpha, \beta > 0$ for the $k+1$ probability levels,

$$\pi^{(j)} \overset{iid}{\sim} \operatorname{Be}(\alpha, \beta), \quad j = 1, \ldots, k+1.$$

The hyperparameter $\boldsymbol{\phi}$ has elements $\alpha, \beta$ here, and $f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{\phi}) = \operatorname{Be}(\pi^{(j)} \,|\, \alpha, \beta)$.

For the block marginal likelihood (3.2.6) we have

$$
\begin{aligned}
f_{block}(\boldsymbol{y}_{\mathcal{S}}) &= \int_0^1 \prod_{t\in\mathcal{S}} \mathrm{Bin}(y_t \,|\, n_t, \pi^{(j)}) \cdot \mathrm{Be}(\pi^{(j)} \,|\, \alpha, \beta)\, d\pi^{(j)} \\
&= \int_0^1 \prod_{t\in\mathcal{S}} \binom{n_t}{y_t} (\pi^{(j)})^{y_t} (1-\pi^{(j)})^{n_t - y_t} \cdot \frac{1}{B(\alpha,\beta)} (\pi^{(j)})^{\alpha-1} (1-\pi^{(j)})^{\beta-1}\, d\pi^{(j)} \\
&= \frac{\prod_{t\in\mathcal{S}} \binom{n_t}{y_t}}{B(\alpha,\beta)} \int_0^1 (\pi^{(j)})^{\sum_{t\in\mathcal{S}} y_t + \alpha - 1} (1-\pi^{(j)})^{\sum_{t\in\mathcal{S}}(n_t - y_t)+\beta-1}\, d\pi^{(j)} \\
&= \frac{\prod_{t\in\mathcal{S}} \binom{n_t}{y_t}}{B(\alpha,\beta)} B\left( \sum_{t\in\mathcal{S}} y_t + \alpha, \sum_{t\in\mathcal{S}} (n_t - y_t) + \beta \right).
\end{aligned}
$$

The derivation of this block density is analogue to the derivation of a binomial-beta density.

#### Posterior

In order to sample the probability parameters given the change points, we need the block posterior density (3.2.7):

$$
\begin{aligned}
f_{block}(\pi^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}, \alpha, \beta) &\propto f(\boldsymbol{y}_{\mathcal{S}} \,|\, y_s, s \in \mathcal{S}, \text{ share the same parameter } \pi^{(j)}) f(\pi^{(j)} \,|\, \alpha, \beta) \\
&= \prod_{t\in\mathcal{S}} \mathrm{Bin}(y_t \,|\, n_t, \pi^{(j)}) \cdot \mathrm{Be}(\pi^{(j)} \,|\, \alpha, \beta) \\
&\propto (\pi^{(j)})^{\sum y_{t\in\mathcal{S}} + \alpha - 1} (1-\pi^{(j)})^{\sum_{t\in\mathcal{S}}(n_t - y_t)+\beta-1} \\
&\propto \mathrm{Be}\left( \pi^{(j)} \,|\, \sum_{t\in\mathcal{S}} y_t + \alpha, \sum_{t\in\mathcal{S}} (n_t - y_t) + \beta \right).
\end{aligned}
$$

### 3.5.2 Case study

We use the data set on rainfall in Tokyo for the years 1983 and 1984, which has been introduced by Kitagawa (1987, p. 1039) as an example of a nonstationary binary process. The data set gives information for all $n = 366$ days if it rained neither in 1983 nor 1984 on this day in Tokyo, in only one of both years, or in both years. All calendar days were passed twice, except day number 60, which is the 29th February of the leap year 1984 and was thus only was passed once (there was no rain). The time series of the relative rain frequencies is plotted in Figure 3.17.

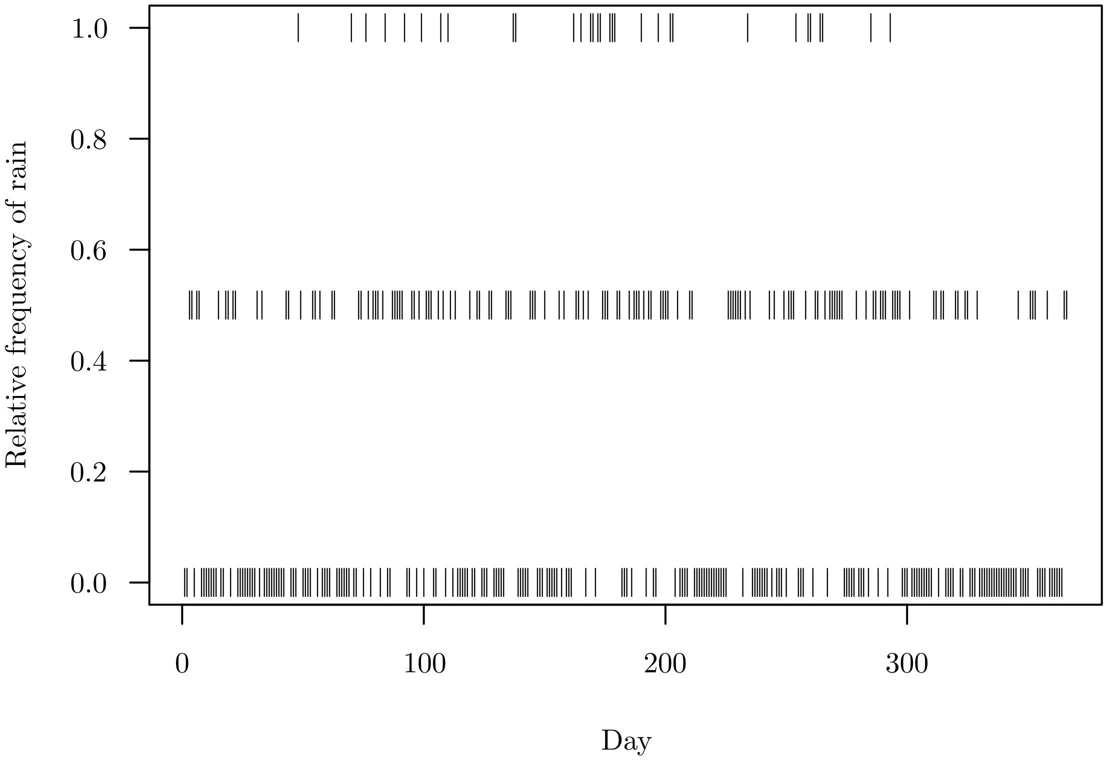


***Figure 3.17*** – *Tokyo rainfall data: Relative frequency of rain in Tokyo per calendar day, over the years 1983 and 1984.*

### Model fitting

We make the assumption that the probability $\pi_t$ of rainfall is constant over the years for each calendar day $t = 1, 2, \dots, 366$. Moreover, we assume that the binary rainfall events are independent conditional on the probabilities. Thus we arrive at the Binomial model, where the number of Bernoulli trials is $n_t = 2$ for $t \neq 60$ and $n_{60} = 1$, and the response $y_t$ is the count of rainy calendar days $t$ during 1983 and 1984.

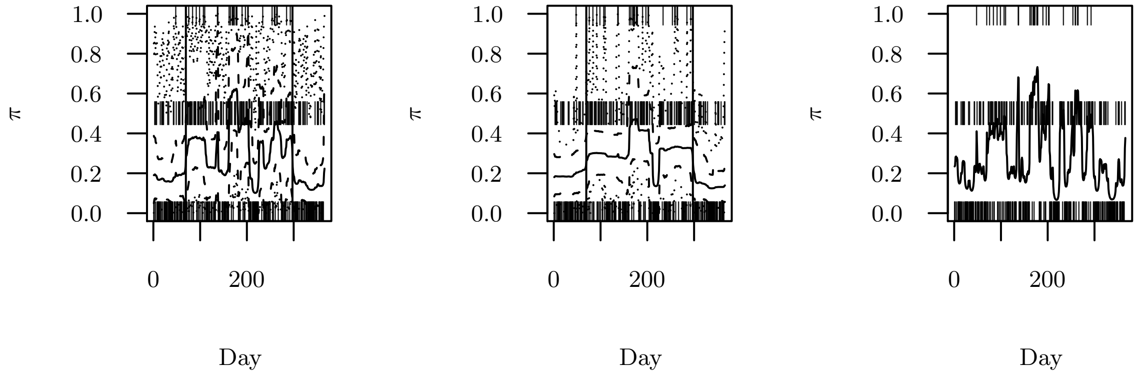
The first model we will fit to the data uses the flat number prior for the change points, and hyperparameters $\alpha = 1, \beta = 1$ for the probabilities prior. This corresponds to a uniform distribution with prior mean and variance

$$\begin{aligned}\mathbb{E}(\pi^{(j)}) &= \frac{\alpha}{\alpha+\beta} = \frac{1}{2},\\ \operatorname{Var}(\pi^{(j)}) &= \mathbb{E}(\pi^{(j)}) \cdot \frac{\beta}{(\alpha+\beta)(\alpha+\beta+1)} = \frac{1}{12} = 0.083.\end{aligned}$$

The second model we want to assess also uses the flat number prior for the change points, but with hyperparameters $\alpha = 0.1, \beta = 0.1$ for the probabilities prior. So the prior mean of the probabilities still equals 1/2, but the variance is now larger at $5/24 = 0.208$, leading to a vaguer prior.

The last model we consider uses the binomial number prior with probability $\pi = 0.2$ for a change point between any two years of the time series. The probabilities prior hyperparameters are chosen as for the first model.

***Figure 3.18*** – *Posterior probabilities trends for the three change point models. Pointwise HPD (dashed lines) as well as simultaneous (dotted lines) 95% credible intervals, which were estimated by simulating 10 000 samples, for the probabilities trend are given. The change point locations in the respective MAP models are marked with vertical lines. (Intervals and MAP change point locations have been omitted for clarity for model 3.)*



**(a)** Model 1

**(b)** Model 2

**(c)** Model 3 mean trend with original data

We have produced 10 000 samples each from the posterior distributions. The estimated probabilities trends and the change point locations in the MAP model are shown in Figure 3.18. The two models with the flat change points prior are similar: Both model 1 in panel (a) and model 2 in panel (b) have their MAP model change points at days $t = 69, 297$. The posterior probabilities for these configurations are $9.49 \cdot 10^{-6}$ and $4.88 \cdot 10^{-3}$, respectively. Analogously to the Poisson-Gamma models in section 3.4.2, the posterior probabilities trend averaged over the change point configurations is much more variable for model 1 than for model 2. Yet, model 3 in panel (c) exhibits a very rough probabilities trend: the MAP model has probability $5.45 \cdot 10^{-48}$ and contains 42 change points. Model 3 thus shows symptoms of overfitting.

The log marginal likelihood values $\log f(\boldsymbol{y})$ of the three change point models are $-325.259$, $-335.244$ and $-331.619$, respectively. So if we should decide on the basis of the log marginal likelihood, model 1 would be our best choice. In the following, we make a more thorough predictive assessment of the three models.

### One-step-ahead predictive assessment

Good one-step-ahead prediction for rainy days is very important, and the weather forecasts use a huge amount of meteorological data to arrive at good predictions. In the context of our example, it is rather the climatological perspective in which we are interested, because we are averaging the two observed years.

First, we generate 10 000 probabilities samples, both from the exact and the approximate one-step-ahead predictive distributions, for all three models. That is for each model, and for all last times $t = 0, 1, \dots, n-1 = 365$, we sample 10 000 variates exactly from $f(\pi_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$ and again 10 000 variates from the approximation $\tilde{f}(\pi_{t+1} \,|\, \boldsymbol{y}_{[1,t]})$. Altogether, this takes 2547, 2391, 2511 seconds for the exact sampling and 378, 218, 454 seconds for the approximate sampling, for the three different models, respectively. Note that the computational effort is much higher here than for the shorter time series on page 41, and that the approximate sampling saves relatively more time.

Second, we plug each probability sample $\pi_t$ into the Binomial likelihood and keep one Binomial variate $y_t^* \sim \operatorname{Bin}(\pi_t, n_t)$ as a sample from the (approximated) one-step-ahead predictive distribution $F_t$ for time $t$ given all prior times.

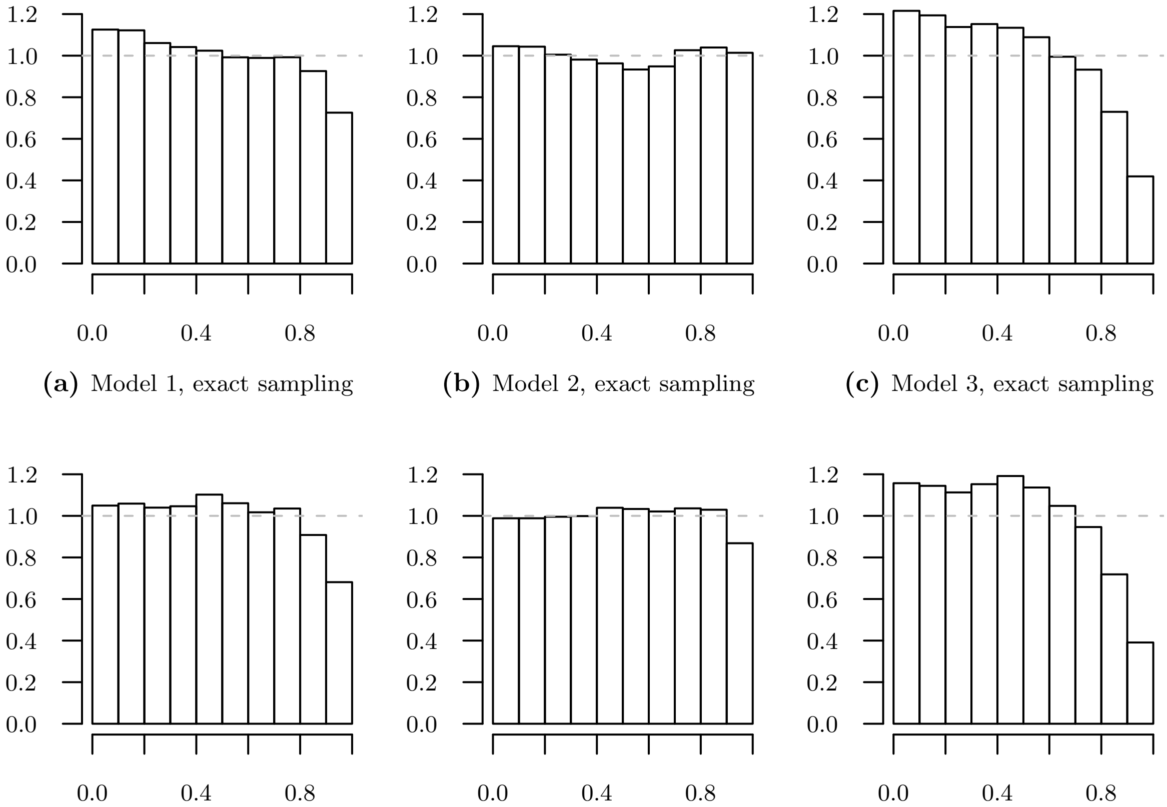
The estimated PIT histograms are shown in Figure 3.19. Only model 2 looks well calibrated, both in the exact histogram (b) and the approximate histogram (e). Model 1 shows a tendency to overestimate the rain probabilities, because the upper histogram bins in panel (a) have too low coverage, while the lower bins have too high coverage. For model 3 in panel (c), this tendency is even more pronounced. The corresponding approximate histograms (d) and (f) essentially match their exact counterparts.

The mid-PIT values are compared between the exact and approximate sampling schemes in Figure 3.20. On the one hand, for the variable model 3 in panel (c), only for a single day a relevant deviation of the approximation is observed. On the other hand, for model 1 in panel (a) and model 2 in panel (b) there are more large differences.

Now we turn to proper scoring rules. The exact and approximate scores of both the ranked probability and the logarithmic scoring rules are compared in Figure 3.21. For model 1, only for a few days at the beginning of the time series there are larger discrepancies. For the log-scores in panel (d) the overall picture is similar to the RPS in panel (a). The same can be said about model 3 in panels (c) and (f), while the absolute deviations of the approximate score values are even smaller (no score approximation is more than 0.5 away from the exact score). Yet, for model 2 in panels (b) and (e), there are more larger deviations than in model 1, and especially the approximations of the logarithmic scores do not work well.

We plot the differences of the approximate and exact mid-PIT values, ranked probability and logarithmic scores versus the day of year in Figure 3.22. The overfitting model 3

***Figure 3.19*** – *PIT histograms for calibration assessment of the one-step-ahead prediction in the three change point models (columns). The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*



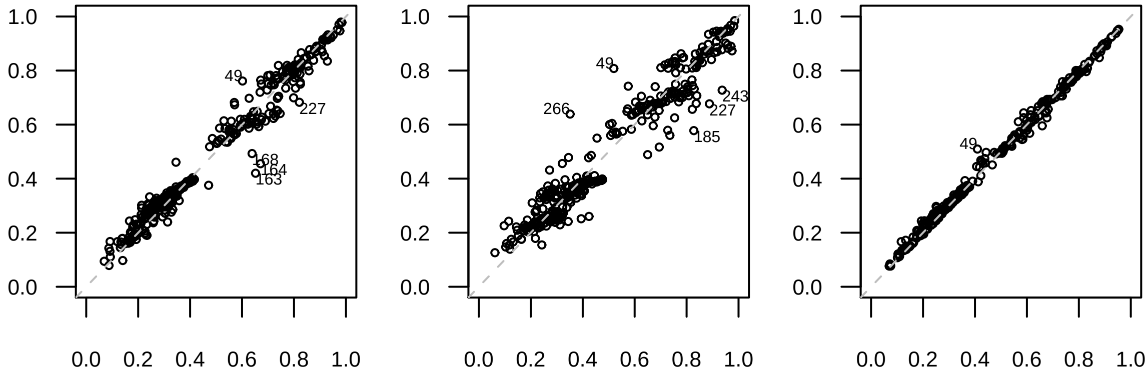
**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

shows smaller differences than both other models, as we have already seen in the previous comparison plots. The parsimonious model 2 features the largest differences, quite surprisingly not only around the estimated probability trend steps from Figure 3.19b, but also between day 250 and 300 and at the end of the time series. The curves for the more variable model 1 mostly follow the model 2 curves with smaller amplitudes.

The mean scores for the proper scoring rules assessment of the one-step-ahead prediction are summarized in Table 3.3. Looking at both RPS rows in the table, it is not surprising that the paired permutation test clearly rejects the hypotheses of same means in the exact and approximate RPS values (p-values $1 \cdot 10^{-4}$ for model 1, $2 \cdot 10^{-4}$ for model 2 and $1 \cdot 10^{-4}$ for model 3). Also if we directly compare the exact and approximate log-scores of each model, the formal test shows the conservativeness of the approximate log-scores. However, model 1 scores highest both under the exact and the approximate sampling scheme, both

***Figure 3.20*** – *Comparison of exact (x-axis) and approximate (y-axis) mid-PIT values for calibration assessment of the one-step-ahead prediction in the three change point models. At most 5 time points where the absolute difference between the two values exceeds 0.1 are labelled.*



(a) Model 1 (b) Model 2 (c) Model 3

considering the RPS and the log-score. Therefore, the model choice using one of these two scoring rules for one-step-ahead predictive assessment would not be changed when the lightweight sampling scheme is used.

***Table 3.3*** – *Mean ranked probability and logarithmic scores for the one-step-ahead prediction of the three models, under the exact and approximate sampling schemes.*

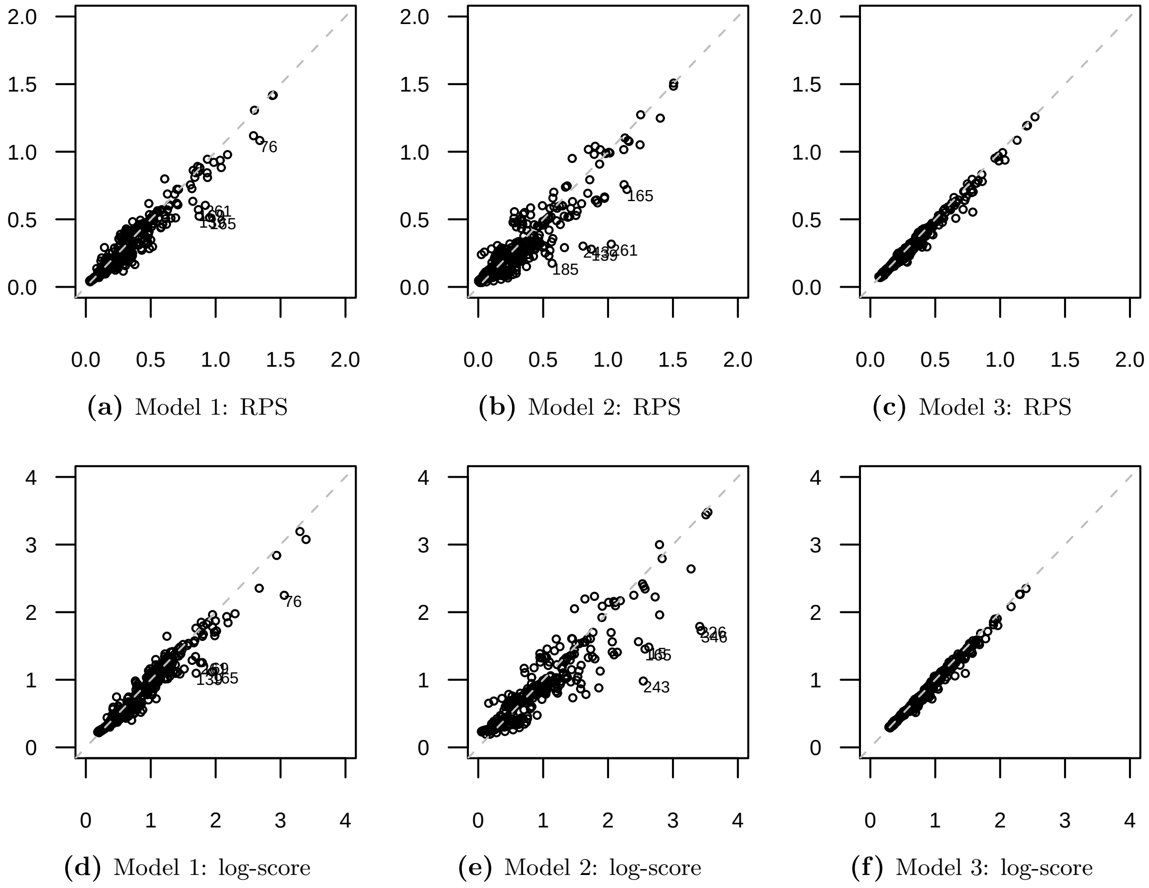
| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| RPS | exact | 0.31 | 0.32 | 0.32 |
| | approximate | 0.29 | 0.30 | 0.30 |
| log-score | exact | 0.89 | 0.92 | 0.91 |
| | approximate | 0.84 | 0.85 | 0.87 |

### Leave-one-out predictive assessment

Next, we will leave out the data from each day in turn, and try predicting it from the remaining data.

First, we generate 10 000 probabilities samples, both from the exact and the approximate leave-one-out distributions, for all three models. Altogether, this takes 8477, 8402, 7904 seconds for the exact sampling and 466, 420, 444 seconds for the approximate sampling, for the three different models, respectively. So the approximate sampling saves more than an order of magnitude of computing time.

***Figure 3.21*** – *Comparison of exact (x-axis) and approximate (y-axis) scores for one-step-ahead prediction in the three change point models (columns). The panels in the upper row compare the RPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact and approximate score values exceeds 0.25 (RPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: RPS

**(b)** Model 2: RPS

**(c)** Model 3: RPS

**(d)** Model 1: log-score

**(e)** Model 2: log-score

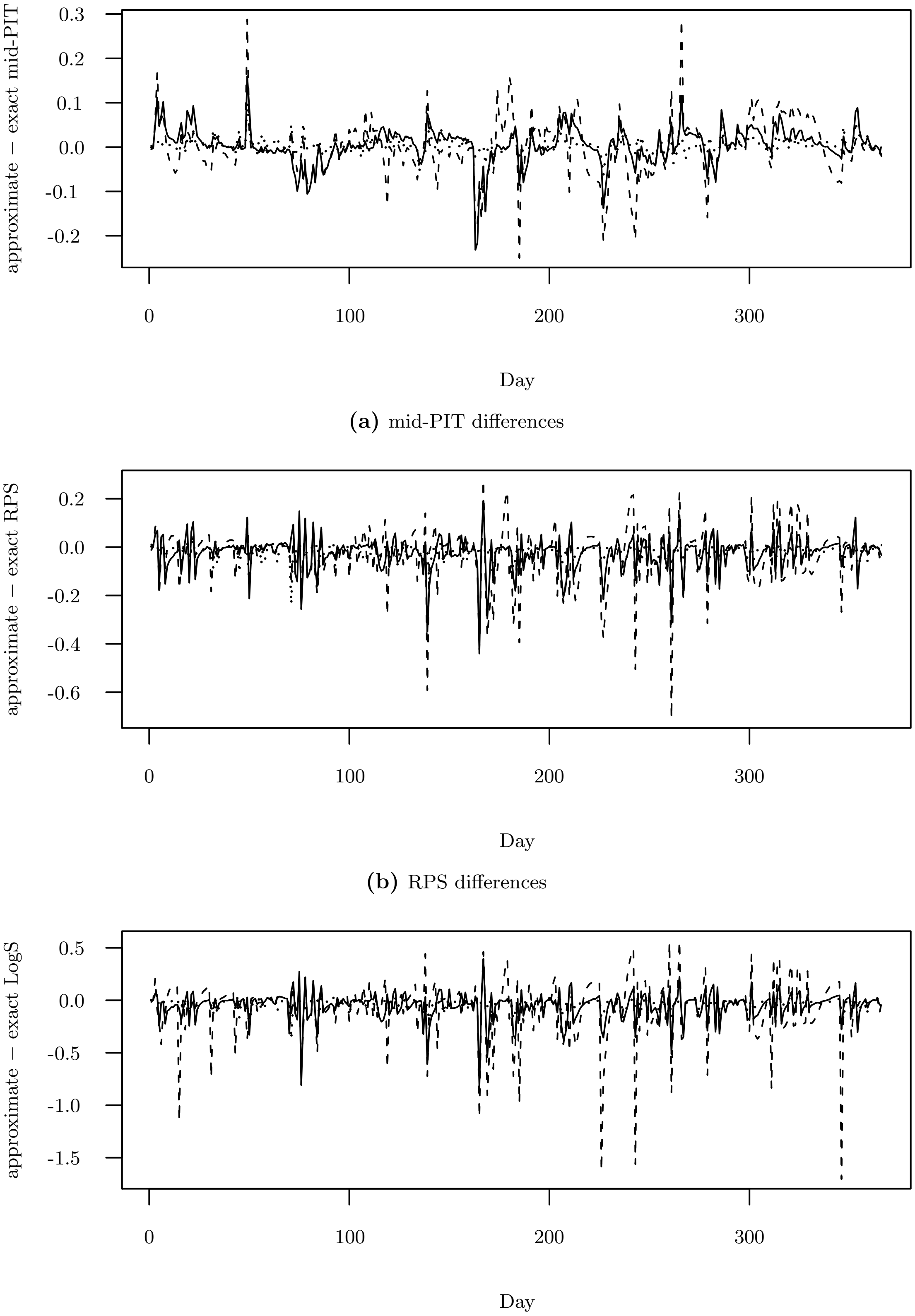
**(f)** Model 3: log-score

Second, we plug each probability sample $\pi_t$ into the Binomial likelihood and keep one Binomial variate $y_t^* \sim \text{Bin}(\pi_t, n_t)$ as a sample from the (approximated) leave-one-out predictive distribution $F_t$ for time $t$ given all other times.

The PIT histograms are shown in Figure 3.23. Model 1 and model 2 look well calibrated if we judge them by panel (a) and panel (b), respectively. The approximate results in panels (d) and (e) are quite similar to their exact counterparts, with a slight tendency to signalling overdispersion for model 1. For model 3 in panel (c) we see again an overestimation picture, which is even more pronounced in the approximate panel (f).

The mid-PIT values are compared between the exact and approximate sampling schemes in Figure 3.24. For model 1 in panel (a), no large deviation of an approximate PIT value from the exact PIT value is noticeable. For model 3 and model 2, only one and two days

***Figure 3.22*** – *Differences of the approximate and exact mid-PIT values, ranked probability and logarithmic scores for the one-step-ahead prediction, for model 1 (——), model 2 (- - -) and model 3 (........).*

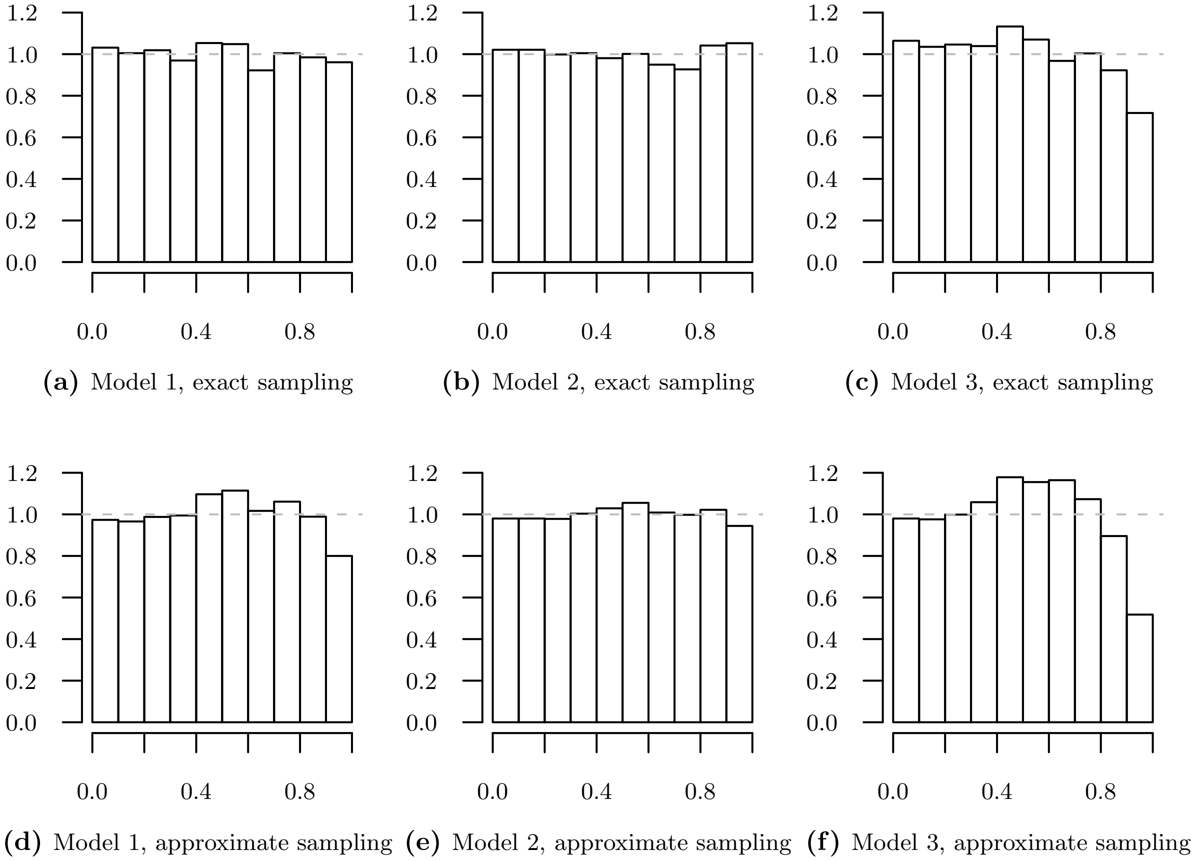


**(a)** mid-PIT differences

**(b)** RPS differences

**(c)** Log-score differences

***Figure 3.23*** – *PIT histograms for calibration assessment of the leave-one-out prediction in the three change point models (columns). The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*



**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

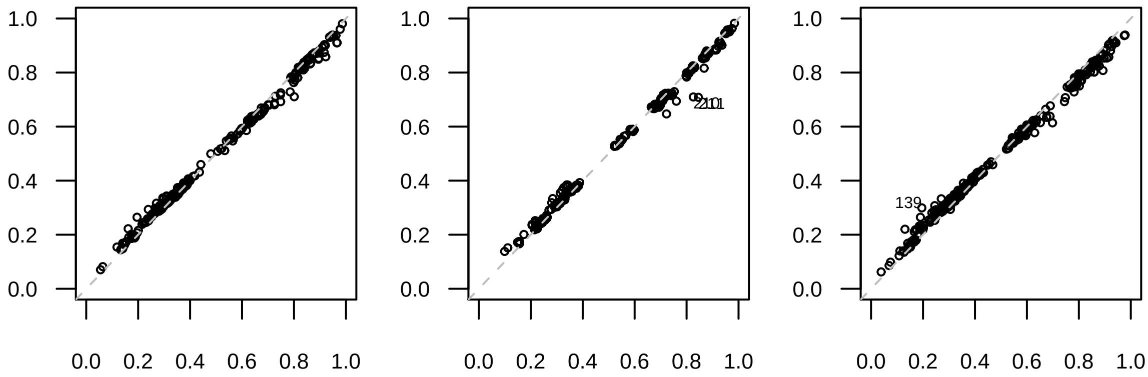
**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

around high jumps in the model-averaged probabilities trend show larger discrepancies between the sampling schemes in panels (c) and (b), respectively.

The exact and approximate scores of both proper scoring rules are compared in Figure 3.25. The approximation is not as good as for the Poisson-Gamma models, cf. Figure 3.13 on page 51. Particularly, it is interesting that here the overfitting model 3 shows more large score discrepancies than in the one-step-ahead assessment and the overfitting Poisson-Gamma model.

The mean scores for the proper scoring rules assessment of the leave-one-out prediction are summarized in Table 3.4 on page 66. These aggregated results are more encouraging than the pairwise comparison of the scores: both for the RPS and the log-score, the ranking of the models (model 3 is best, then model 1 and model 2) is unaltered when we use the approximate scores instead of the exact scores.

***Figure 3.24*** – *Comparison of exact (x-axis) and approximate (y-axis) mid-PIT values for calibration assessment of the leave-one-out prediction in the three change point models. Time points where the absolute difference between the two values exceeds 0.1 are labelled.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

### Posterior-predictive checking

Now we will look at the results of posterior-predictive model checking.

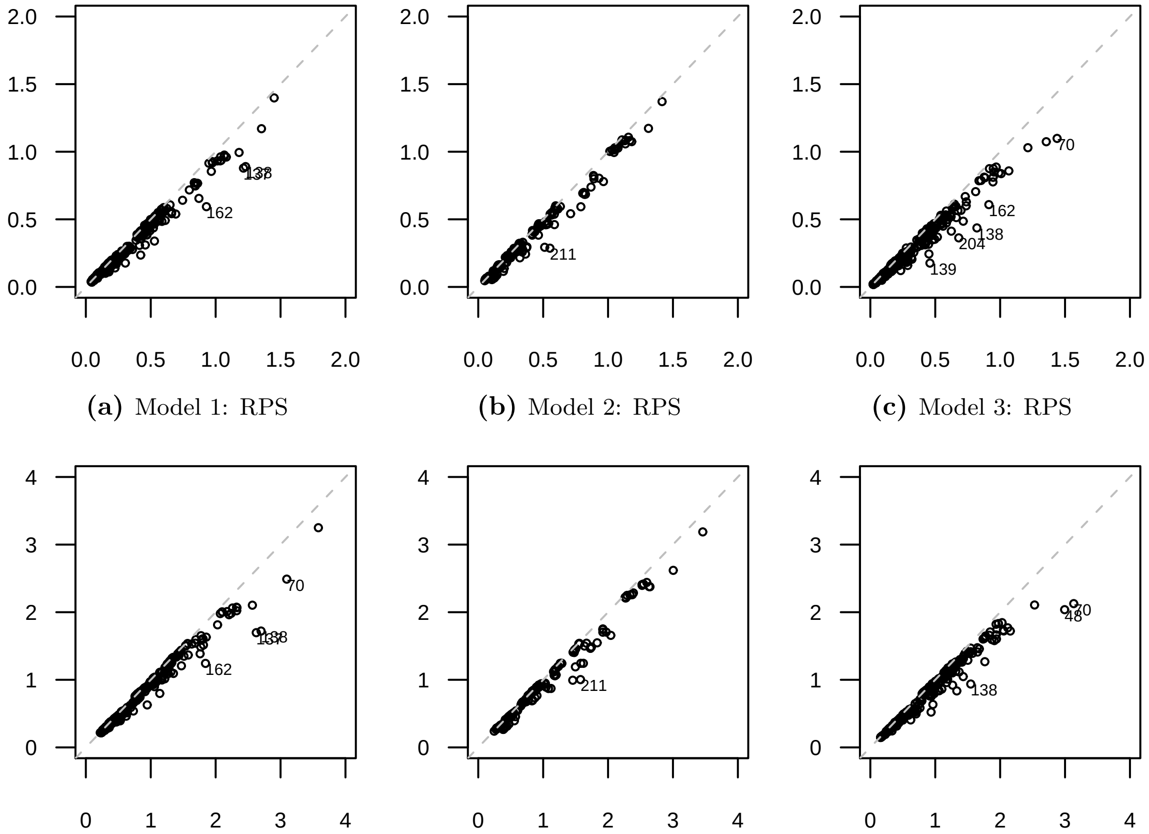
The posterior-predictive PIT histograms are shown in Figure 3.26. While for model 2, panel (b) still shows a good calibration for the leave-one-out prediction, the other two models would be judged differently compared to the exact or approximate leave-one-out assessment. panel (a) argues for a slight overdispersion of model 1, and panel (c) shows an extreme overdispersion of model 3, which is neither apparent in the exact panel (c) nor the approximate panel (f) of Figure 3.23 with the leave-one-out results.

The mid-PIT values are compared between the exact leave-one-out and the posterior-predictive sampling schemes in Figure 3.27. Again, substantial shrinkage of the mid-PIT values towards 0.5 can be seen in the panels, which is strongest for model 3 in panel (c). This explains the stronger overdispersion pictures in Figure 3.26.

The exact leave-one-out scores are compared with the posterior-predictive scores in Figure 3.28. For model 1 and model 2 in panels (a) – (e), the differences to the respective plots comparing the exact with the approximate leave-one-out scores in Figure 3.25 are not very large. Yet, panels (c) and (f) for model 3 show a much worse approximation than the counterparts in Figure 3.25.

The mean scores are summarized and compared to the leave-one-out scores in Table 3.4. The heavy bias of individual model 3 posterior-predictive scores which we observed in Figure 3.28 is mirrored in the corresponding mean RPS and log-scores: if we only looked at the mean posterior-predictive model scores, model 3, the overfitting model with the

***Figure 3.25*** – *Comparison of exact (x-axis) and approximate (y-axis) scores for leave-one-out prediction in the three change point models (columns). The panels in the upper row compare the RPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact and approximate score values exceeds 0.25 (RPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: RPS

**(b)** Model 2: RPS

**(c)** Model 3: RPS

**(d)** Model 1: log-score

**(e)** Model 2: log-score

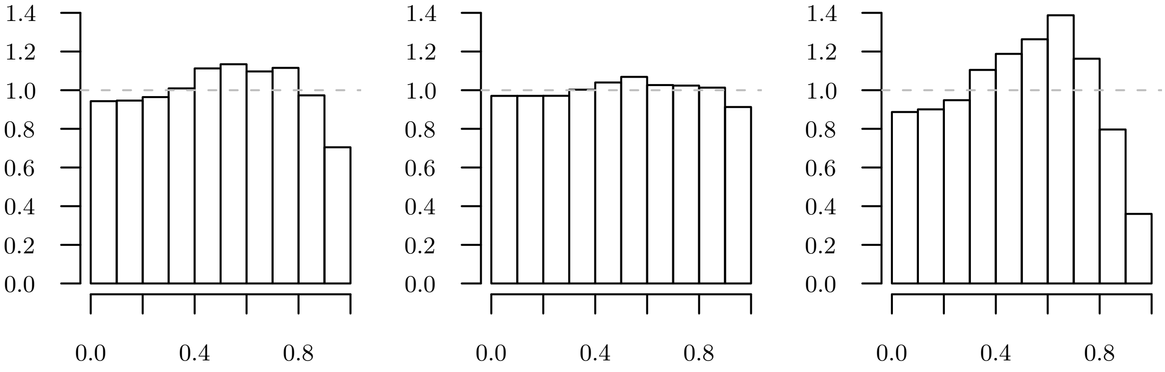
**(f)** Model 3: log-score

most variable fit, appears to be much better than the other two models. Yet, using the exact and also the approximate mean scores, the difference between model 3 and model 1 is not very large.

### Results

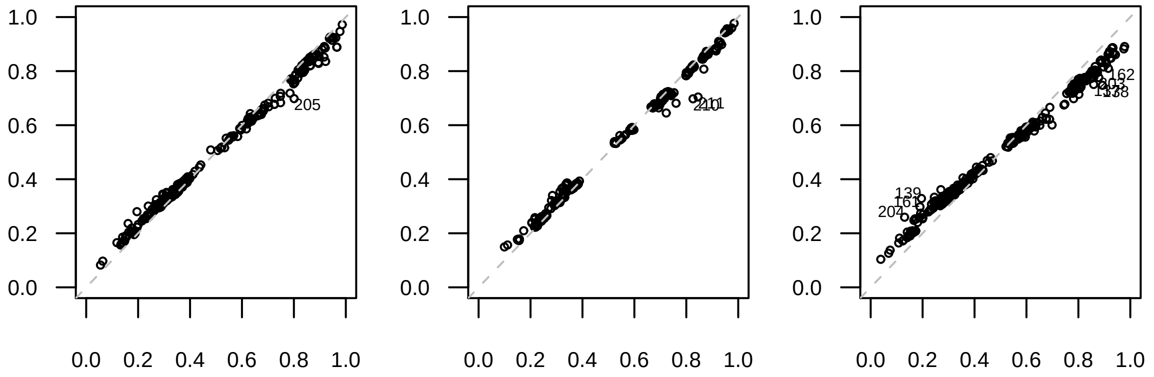
Starting with the exact results, model 1 looks best in the one-step-ahead scores, with an acceptable calibration in the corresponding PIT histogram. In the leave-one-out scores, model 3 gets ahead of model 1, but its PIT histogram is slightly overdispersed. Compared to the marginal likelihood rating, where model 1 was clearly preferred, the leave-one-out assessment might tend to preferring overfitting models.

***Figure 3.26*** *– PIT histograms for posterior-predictive checking the three change point models.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

***Figure 3.27*** *– Comparison of exact leave-one-out (x-axis) and posterior-predictive (y-axis) mid-PIT values in the three change point models. Time points where the absolute difference between the two values exceeds 0.1 are labelled.*

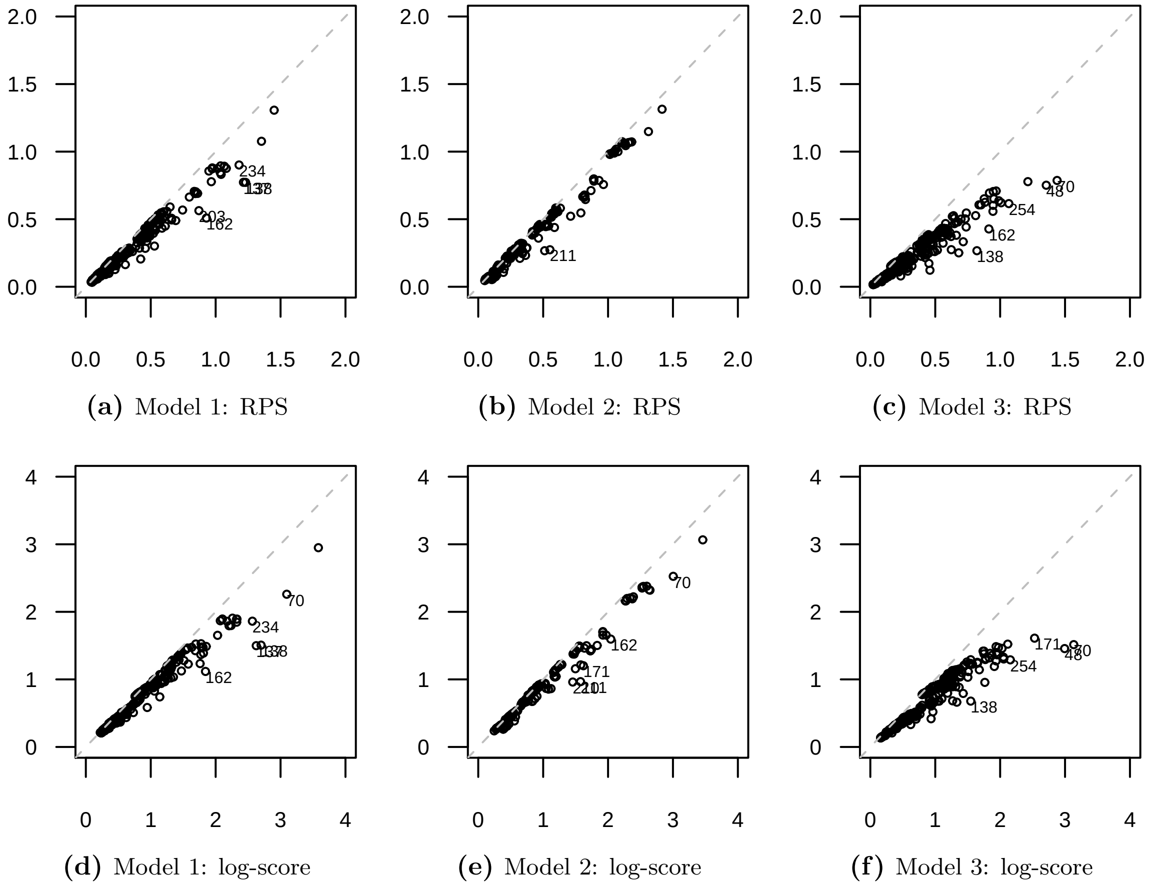


**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

***Table 3.4*** *– Mean ranked probability and logarithmic scores for the three models, under the exact and approximate leave-one-out and the posterior predictive sampling schemes.*

| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| RPS | exact leave-one-out | 0.30 | 0.31 | 0.29 |
| | approximate leave-one-out | 0.27 | 0.29 | 0.25 |
| | posterior-predictive | 0.25 | 0.28 | 0.21 |
| log-score | exact leave-one-out | 0.85 | 0.87 | 0.84 |
| | approximate leave-one-out | 0.79 | 0.82 | 0.76 |
| | posterior-predictive | 0.75 | 0.81 | 0.67 |

***Figure 3.28*** – *Comparison of exact leave-one-out (x-axis) and posterior-predictive (y-axis) scores in the three change point models (columns). The panels in the upper row compare the RPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact leave-one-out and posterior-predictive score values exceeds 0.25 (RPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: RPS **(b)** Model 2: RPS **(c)** Model 3: RPS

**(d)** Model 1: log-score **(e)** Model 2: log-score **(f)** Model 3: log-score

Continuing with the performance of the proposed approximate sampling, it is encouraging that the exact one-step-ahead scores ranking could be replicated as well as the corresponding PIT histograms. The same can be said about the leave-one-out assessment. Especially for this long time series, the gain is computational efficiency is worthwhile.

The posterior-predictive scores signal that model 3 fits the given data best, and produce the same model ranking as the leave-one-out scores. However, the posterior-predictive PIT histograms cannot be used as an approximation to the exact leave-one-out PIT histograms.

## 3.6 Normal-Normal-Gamma model

The specializations of the general framework from section 3.2 which are necessary for the Normal-Normal-Gamma change point model are described in section 3.6.1. Three instances of this model class are fitted to the Nile discharge data and are then subject of predictive assessment in a case study in section 3.6.2.

### 3.6.1 The special change point model

#### Data

The Normal-Normal-Gamma change point model will be suited to modelling of time series $\boldsymbol{y} := (y_1, y_2, \dots, y_n)$ of real-valued observations $y_t \in \mathbb{R}$. Even if the observations are actually restricted to a subset of $\mathbb{R}$, the model can be used if a normal approximation is sensible. For example, in the case study we will model positive-valued discharge levels, but the range of the observations is far enough away from zero to justify the use of the proposed model for real-valued observations.

For this model no covariates $\boldsymbol{x}_t$ are considered. However, an extension to integrate covariates via the conjugate Bayesian linear model would be straightforward.

#### Model

We assume independent normal distributions with means $\mu_t$ and precisions $\kappa_t$ for the observations:

$$y_t \,|\, \mu_t, \kappa_t \overset{ind}{\sim} \mathrm{N}(\mu_t, 1/\kappa_t), \quad t \in \mathcal{N}.$$

The parametrization with the precision instead of the variance is chosen for notational convenience.

So the parameters have two elements for this model, $\boldsymbol{\xi}_t = (\mu_t, \kappa_t)$. The response density is $f(y_t \,|\, \boldsymbol{\xi}_t) = \mathrm{N}(y_t \,|\, \mu_t, 1/\kappa_t)$.

#### Prior

The normal-gamma distribution is conjugate to the normal likelihood when both mean and precision are unknown. So for the $k+1$ parameter levels independent identical normal-gamma priors with hyperparameters $\nu, \lambda, \alpha, \beta > 0$ are specified,

$$(\mu^{(j)}, \kappa^{(j)}) \overset{iid}{\sim} \mathrm{NG}(\nu, \lambda, \alpha, \beta), \quad j = 1, \dots, k+1. \tag{3.6.1}$$

The normal-gamma distribution means that if $(\mu, \kappa) \sim \mathrm{NG}(\nu, \lambda, \alpha, \beta)$, then

$$\begin{aligned} \mu \,|\, \kappa &\sim \mathrm{N}(\nu, (\lambda\kappa)^{-1}) \\ \text{and} \quad \kappa &\sim \mathrm{G}(\alpha, \beta). \end{aligned}$$

The hyperparameter $\boldsymbol{\phi}$ has thus four elements $\nu, \lambda, \alpha, \beta$ here, and the parameter level prior density is $f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{\phi}) = \mathrm{NG}(\mu^{(j)}, \kappa^{(j)} \,|\, \nu, \lambda, \alpha, \beta)$.

Plugging in the block posterior $\mu^{(j)}, \kappa^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}} \sim \mathrm{NG}(\nu_{\mathcal{S}}, \lambda_{\mathcal{S}}, \alpha_{\mathcal{S}}, \beta_{\mathcal{S}})$ from (3.6.2), we have from (3.2.6) that the block marginal likelihood is

$$\begin{aligned}
f_{block}(\boldsymbol{y}_{\mathcal{S}}) &= \frac{\prod_{t\in\mathcal{S}} f(y_t \,|\, \boldsymbol{\xi}^{(j)}) f(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{\phi})}{f_{block}(\boldsymbol{\xi}^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}, \boldsymbol{\phi})} \\
&= \frac{\prod_{t\in\mathcal{S}} \mathrm{N}(y_t \,|\, \mu^{(j)}, 1/\kappa^{(j)})\, \mathrm{NG}(\mu^{(j)}, \kappa^{(j)} \,|\, \nu, \lambda, \alpha, \beta)}{\mathrm{NG}(\mu^{(j)}, \kappa^{(j)} \,|\, \nu_{\mathcal{S}}, \lambda_{\mathcal{S}}, \alpha_{\mathcal{S}}, \beta_{\mathcal{S}})} \\
&= (2\pi)^{-\frac{n_{\mathcal{S}}}{2}} (\kappa^{(j)})^{\frac{n_{\mathcal{S}}}{2}} \exp\left( -\frac{\kappa^{(j)}}{2} \sum_{t\in\mathcal{S}} (y_t - \mu^{(j)})^2 \right) \\
&\quad \times \frac{(2\pi)^{-\frac{1}{2}} (\lambda\kappa^{(j)})^{\frac{1}{2}} \exp(-\frac{\lambda\kappa^{(j)}}{2}(\mu^{(j)} - \nu)^2) \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\kappa^{(j)})^{\alpha-1} \exp(-\kappa^{(j)}\beta)}{(2\pi)^{-\frac{1}{2}} (\lambda_{\mathcal{S}}\kappa^{(j)})^{\frac{1}{2}} \exp(-\frac{\lambda_{\mathcal{S}}\kappa^{(j)}}{2}(\mu^{(j)} - \nu_{\mathcal{S}})^2) \frac{(\beta_{\mathcal{S}})^{\alpha_{\mathcal{S}}}}{\Gamma(\alpha_{\mathcal{S}})} (\kappa^{(j)})^{\alpha_{\mathcal{S}}-1} \exp(-\kappa^{(j)}\beta_{\mathcal{S}})} \\
&= \left( \frac{\lambda}{\lambda_{\mathcal{S}}} \right)^{\frac{1}{2}} \frac{\beta^{\alpha}}{(\beta_{\mathcal{S}})^{\alpha_{\mathcal{S}}}} \frac{\Gamma(\alpha_{\mathcal{S}})}{\Gamma(\alpha)} (2\pi)^{-\frac{n_{\mathcal{S}}}{2}} \\
&= \left( \frac{\lambda}{\pi^{n_{\mathcal{S}}}(n_{\mathcal{S}} + \lambda)} \right)^{\frac{1}{2}} \frac{\Gamma(\frac{n_{\mathcal{S}}}{2} + \alpha)}{\Gamma(\alpha)} (2\beta)^{\alpha} \\
&\quad \times \left\{ n_{\mathcal{S}}\hat{v}_{\mathcal{S}} + \frac{\lambda n_{\mathcal{S}}}{n_{\mathcal{S}} + \lambda} (\hat{m}_{\mathcal{S}} - \nu)^2 + 2\beta \right\}^{-(\frac{n_{\mathcal{S}}}{2} + \alpha)},
\end{aligned}$$

where $n_{\mathcal{S}} := |\mathcal{S}|$ denotes the number of time points in the set $\mathcal{S}$, while $\hat{m}_{\mathcal{S}} := \frac{1}{n_{\mathcal{S}}} \sum_{t\in\mathcal{S}} y_t$ and $\hat{v}_{\mathcal{S}} := \frac{1}{n_{\mathcal{S}}} \sum_{t\in\mathcal{S}} (y_t - \hat{m}_{\mathcal{S}})^2$ abbreviate the empirical mean and variance, respectively, in the observations $\boldsymbol{y}_{\mathcal{S}}$.

### Posterior

In order to sample the model parameters given the change points, we need the block posterior density (3.2.7):

$$
\begin{aligned}
&f_{block}(\mu^{(j)}, \kappa^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}) \\
&\qquad \propto \prod_{t\in\mathcal{S}} f(y_t \,|\, \mu^{(j)}, \kappa^{(j)}) f(\mu^{(j)}, \kappa^{(j)}) \\
&\qquad = \prod_{t\in\mathcal{S}} \mathrm{N}(y_t \,|\, \mu^{(j)}, 1/\kappa^{(j)}) \, \mathrm{NG}(\mu^{(j)}, \kappa^{(j)} \,|\, \nu, \lambda, \alpha, \beta) \\
&\qquad = (2\pi)^{-\frac{n_{\mathcal{S}}}{2}} (\kappa^{(j)})^{\frac{n_{\mathcal{S}}}{2}} \exp\left(-\frac{\kappa^{(j)}}{2}\sum_{t\in\mathcal{S}}(y_t-\mu^{(j)})^2\right) \\
&\qquad\quad \times (2\pi)^{-\frac{1}{2}} (\lambda\kappa^{(j)})^{\frac{1}{2}} \exp\left(-\frac{\lambda\kappa^{(j)}}{2}(\mu^{(j)}-\nu)^2\right) \frac{\beta^\alpha}{\Gamma(\alpha)} (\kappa^{(j)})^{\alpha-1} \exp(-\kappa^{(j)}\beta) \\
&\qquad \propto (\kappa^{(j)})^{\frac{n_{\mathcal{S}}+1}{2}+\alpha-1} \exp\left\{-\kappa^{(j)}\left[\frac{1}{2}\sum_{t\in\mathcal{S}}(y_t-\mu^{(j)})^2 + \frac{\lambda}{2}(\mu^{(j)}-\nu)^2 + \beta\right]\right\}.
\end{aligned}
$$

Now the term in square brackets can be rewritten as a quadratic form in $\mu^{(j)}$, which gives us the normal-gamma shape:

$$
\begin{aligned}
&f_{block}(\mu^{(j)}, \kappa^{(j)} \,|\, \boldsymbol{y}_{\mathcal{S}}) \\
&\quad \propto (\kappa^{(j)})^{\frac{n_{\mathcal{S}}+1}{2}+\alpha-1} \\
&\qquad \times \exp\left\{-\kappa^{(j)}\left[\frac{n_{\mathcal{S}}}{2}\hat{v}_{\mathcal{S}} + \frac{1}{2}\frac{\lambda n_{\mathcal{S}}}{n_{\mathcal{S}}+\lambda}(\hat{m}_{\mathcal{S}}-\nu)^2 + \beta + \frac{n_{\mathcal{S}}+\lambda}{2}\left(\mu^{(j)} - \frac{n_{\mathcal{S}}\hat{m}_{\mathcal{S}}+\lambda\nu}{n_{\mathcal{S}}+\lambda}\right)^2\right]\right\} \\
&\quad = (\kappa^{(j)})^{\frac{1}{2}} \exp\left\{-\frac{\kappa^{(j)}(n_{\mathcal{S}}+\lambda)}{2}\left(\mu^{(j)} - \frac{n_{\mathcal{S}}\hat{m}_{\mathcal{S}}+\lambda\nu}{n_{\mathcal{S}}+\lambda}\right)^2\right\} \\
&\qquad \times (\kappa^{(j)})^{\frac{n_{\mathcal{S}}}{2}+\alpha-1} \exp\left\{-\kappa^{(j)}\left[\frac{n_{\mathcal{S}}}{2}\hat{v}_{\mathcal{S}} + \frac{1}{2}\frac{\lambda n_{\mathcal{S}}}{n_{\mathcal{S}}+\lambda}(\hat{m}_{\mathcal{S}}-\nu)^2 + \beta\right]\right\} \\
&\quad \propto \mathrm{NG}\left(\mu^{(j)}, \kappa^{(j)} \,|\, \nu_{\mathcal{S}}, \lambda_{\mathcal{S}}, \alpha_{\mathcal{S}}, \beta_{\mathcal{S}}\right), \qquad (3.6.2)
\end{aligned}
$$

where the posterior parameters are

$$
\begin{aligned}
\nu_{\mathcal{S}} &:= \frac{n_{\mathcal{S}}\hat{m}_{\mathcal{S}}+\lambda\nu}{n_{\mathcal{S}}+\lambda}, \\
\lambda_{\mathcal{S}} &:= n_{\mathcal{S}}+\lambda, \\
\alpha_{\mathcal{S}} &:= \frac{n_{\mathcal{S}}}{2}+\alpha \\
\text{and} \quad \beta_{\mathcal{S}} &:= \frac{n_{\mathcal{S}}}{2}\hat{v}_{\mathcal{S}} + \frac{1}{2}\frac{\lambda n_{\mathcal{S}}}{n_{\mathcal{S}}+\lambda}(\hat{m}_{\mathcal{S}}-\nu)^2 + \beta.
\end{aligned}
$$

### 3.6.2 Case study

We illustrate the predictive assessment with the Nile data from Cobb (1978, p. 248). The data set comprises a total of $n = 100$ contiguous yearly discharge measurements of the Nile at Aswan, from 1871 to 1970. The time series is plotted in Figure 3.29. Cobb (1978) assumed the values to be normally distributed conditional on the means, and used conditional inference techniques to search for a single change point in the means after fixing two possible mean values and the variance. We will allow an arbitrary number of change points in the parameters, and assume the mean and variance of the normal distributions in the blocks as unknown.

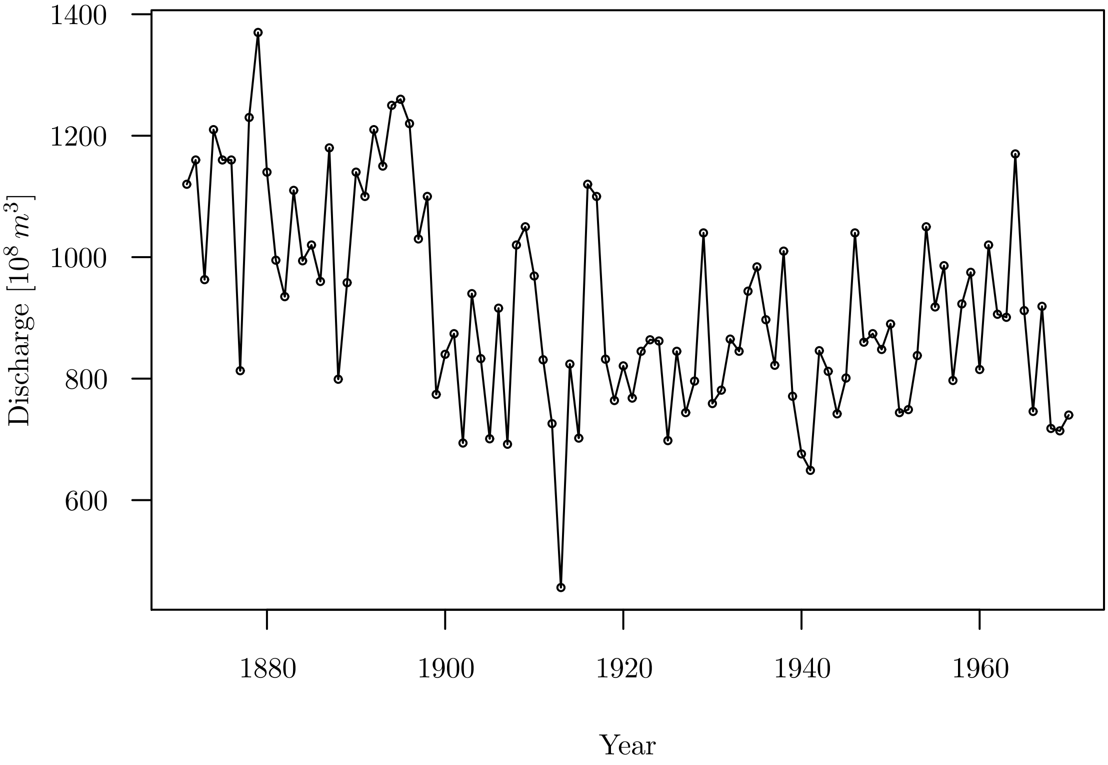


***Figure 3.29*** – *Nile discharge data: yearly discharge levels in* $10^8\,m^3$ *measured at Aswan from 1871 to 1970.*

### Model fitting

We make the assumption that the mean $\mu_t$ and precision $\kappa_t$ of Nile discharge $y_t$ are piecewise constant, and fit change point models with the parameter $\boldsymbol{\xi}_t = (\mu_t, \kappa_t)$. For the prior normal-gamma distribution (3.6.1) of $\boldsymbol{\xi}_t$, we center the prior distribution of $\mu_t$ around the data mean by setting $\nu = 919.35$. The precision factor $\lambda$ is varied between the models. Following Cobb (1978), who fixed the standard deviation at $\sigma = 125$, we choose such gamma distribution parameters for $\kappa_t$ that the prior mean for the variance $\sigma_t^2 = 1/\kappa_t$ is $125^2 = 15\,625$. Since $\sigma_t^2$ is *a priori* inverse gamma distributed with expectation

$\mathbb{E}(\sigma^2) = \beta/(\alpha - 1)$, $\alpha$ and $\beta$ can be chosen appropriately.

The first model we will fit to the data uses the flat number prior for the change points, and hyperparameters $\lambda = 0.1, \alpha = 100 + 1, \beta = 15\,625 \cdot 100$ for the parameters prior.

The second model we want to assess also uses the flat number prior for the change points, but with hyperparameters $\lambda = 0.001, \alpha = 10 + 1, \beta = 15\,625 \cdot 10$ for the parameters prior. So the prior expectations of the mean and variance levels are unaltered, but the prior variances are enlarged. For example, we now have $\operatorname{Var}(\sigma^2) = \mathbb{E}(\sigma^2)^2/(\alpha - 2) = 125^4/9$, compared to $125^4/99$ in the model 1 setting.

The last model we consider uses the binomial number prior with probability $\pi = 0.2$ for a change point between any two years of the time series. The parameters prior hyperparameters are chosen as for the first model.

We have produced 10 000 samples each from the posterior distributions. The estimated parameters trends and the change point locations in the MAP model are shown in Figure 3.30.

The two models with the flat change points prior are similar: Both model 1 in panel (a) and model 2 in panel (b) have one MAP model change point after the year 1898. The posterior probabilities for these configurations are $4.65 \cdot 10^{-1}$ and $7.46 \cdot 10^{-1}$, respectively. While the simultaneous credible band shows a higher variability of the model 1 means in panel (a), the model averaged mean trend is almost indiscernible from the model 2 trend in panel (b): both trends show a clear step downwards around their change points, and are constant elsewhere. The corresponding averaged precision seems to step upwards, more for model 2 than for model 1. Model 3 with the binomial change points prior in panel (c) exhibits a more variable mean trend, which looks overfitted to the data. The MAP model here has probability $3.22 \cdot 10^{-5}$ and contains an additional change point after the year 1967.

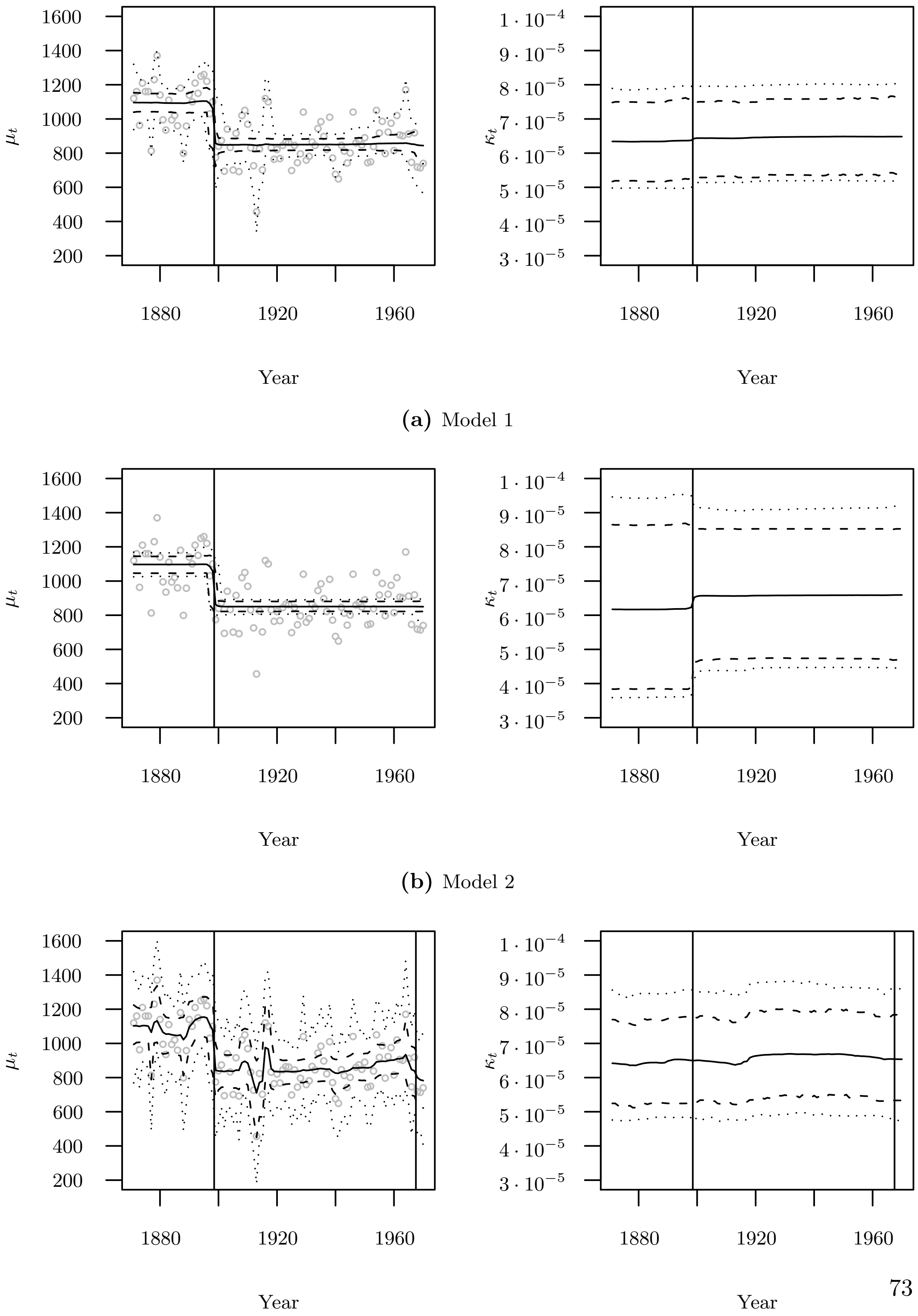
The log marginal likelihood values $\log f(\boldsymbol{y})$ of the three change point models are $-640.72$, $-646.668$ and $-647.005$, respectively. So if we should decide on the basis of the marginal likelihood, model 1 would be our best choice. Whether this choice is supported by a predictive model assessment will be examined in the following.

### One-step-ahead predictive assessment

First, we will do a one-step-ahead predictive assessment of the three models, and compare the approximate results with the exact results.

First, we generate 10 000 parameters samples, both from the exact and the approximate one-step-ahead predictive distributions, for all three models. That is for each model, and for all last times $t = 0, 1, \ldots, n - 1 = 99$, we sample 10 000 variates exactly from $f(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{y}_{[1,t]})$ and again 10 000 variates from the approximation $\tilde{f}(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{y}_{[1,t]})$. Altogether,

***Figure 3.30*** – *Posterior parameters trends for the three change point models. Pointwise HPD (dashed lines) as well as simultaneous (dotted lines) 95% credible intervals, which were estimated by simulating 10 000 samples, for both the mean (left panels) and the precision trends (right panels) are given. The change point locations in the respective MAP models are marked with vertical lines.*



**(a)** Model 1

**(b)** Model 2

**(c)** Model 3

this takes 84, 55, 160 seconds for the exact sampling and 26, 16, 119 seconds for the approximate sampling, for the three different models, respectively. So for model 3, the relative gain in computing time of the approximate sampling approach is rather small. This is probably due to the fact that the computational effort for this short time series lies mainly in the parameter levels sampling, and not in the change points sampling. For the wiggly model 3, more parameter levels need to be sampled than for the other two smooth models.

Second, we plug each parameter sample $\boldsymbol{\xi}_t = (\mu_t, \kappa_t)$ into the normal likelihood and keep one Gaussian variate $y_t^* \sim \mathrm{N}(\mu_t, 1/\kappa_t)$ as a sample from the (approximated) one-step-ahead predictive distribution $F_t$ for time $t$ given all prior times.

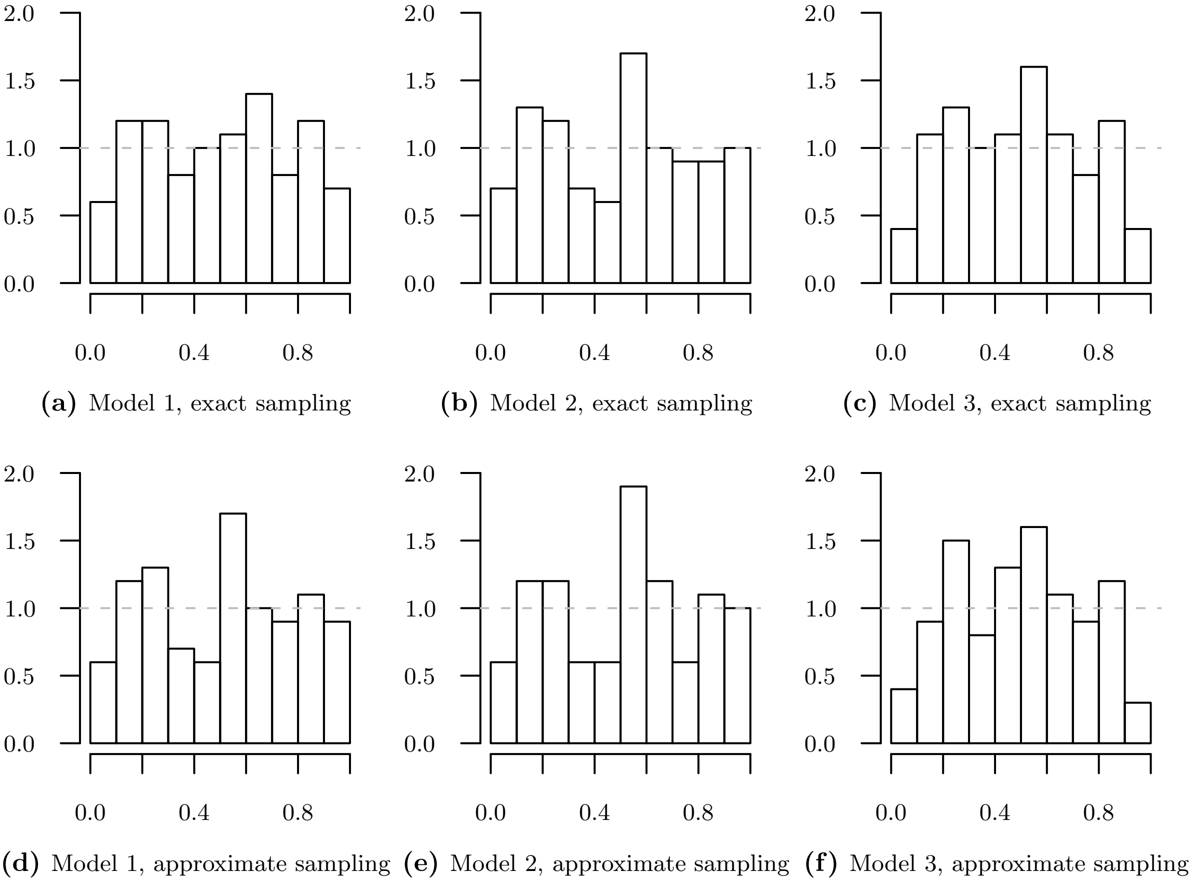
The estimated PIT histograms are shown in Figure 3.31. All exact PIT histograms do not look very good, but the reason could be the relatively small sample size ($n = 100$). The histogram in panel (b) for model 2 signals an acceptable calibration. The approximate histogram in panel (e) does not look as good, but is similar. Also the exact histogram for model 1 in panel (a) shows some differences between nominal and observed prediction intervals coverages, which are still present in the approximate histogram in panel (d). One might diagnose an overdispersion of model 3 from the exact panel (c) and also from the approximate panel (f).

The PIT values are compared between the exact and approximate sampling schemes in Figure 3.32. On the one hand, for the variable model 3, only for a single year a deviation of the approximation larger than 0.1 is observed in panel (c). For model 2 in panel (b), a few years after the MAP change point have larger deviations. On the other hand, for model 1 in panel (a) there are more differences, which are mostly after the change point year.

Now we turn to proper scoring rules. The exact and approximate scores of both the continuous ranked probability and the logarithmic scoring rules are compared in Figure 3.33. Overall, the approximate sampling works well for this example. The most large differences are observed for model 1, both for the CRPS in panel (a) and for the log-score in panel (d). However, it is promising that the points in the figures are distributed quite evenly around the identity line, and do not always lie in the lower-right triangular, which would mean that the approximate score values underestimate the exact score values systematically. There are fewer differences for model 2, where both for the CRPS in panel (b) and for the log-score in (e), the score values for the three years immediately after the MAP model change point 1899 (the three years in the new MAP block) are heavily underestimated by the approximate sampling scheme. For model 3, some larger differences occur for the CRPS in panel (c), while the differences for the log-score in panel (f) are minor.

We plot the time series of differences of the approximate and exact PIT values, con-

***Figure 3.31*** – *PIT histograms for calibration assessment of the one-step-ahead prediction in the three change point models (columns). The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*



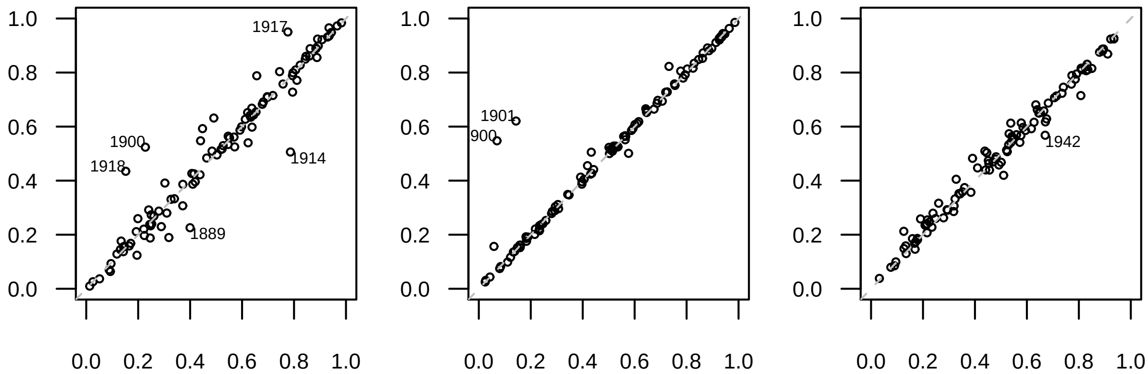
**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

tinuous ranked probability and logarithmic scores in Figure 3.34. Model 1 and model 2 are too optimistic about their forecast performance around the turn of the century, with too low score values in panels (b) and (c). The exact one-step-ahead sampling predicts still high discharge levels, while the observations materialize on a lower level, leading to small PIT values. The approximate sampling knows about the step, and thus produces too large PIT values around 1900, as panel (a) shows. Larger approximation errors are also observed between 1910 and 1920, when the discharge levels fluctuate more (cf. Figure 3.29). Overall the differences seem to diminish in the late years, which is expected because more of the data used by the approximate sampling scheme is also used by the exact sampling scheme.

The mean scores for the proper scoring rules assessment of the one-step-ahead prediction are summarized in Table 3.5. Looking at both CRPS rows in the table, it is not surprising

***Figure 3.32*** – *Comparison of exact (x-axis) and approximate (y-axis) PIT values for calibration assessment of the one-step-ahead prediction in the three change point models. At most 5 time points where the absolute difference between the two values exceeds 0.1 are labelled.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

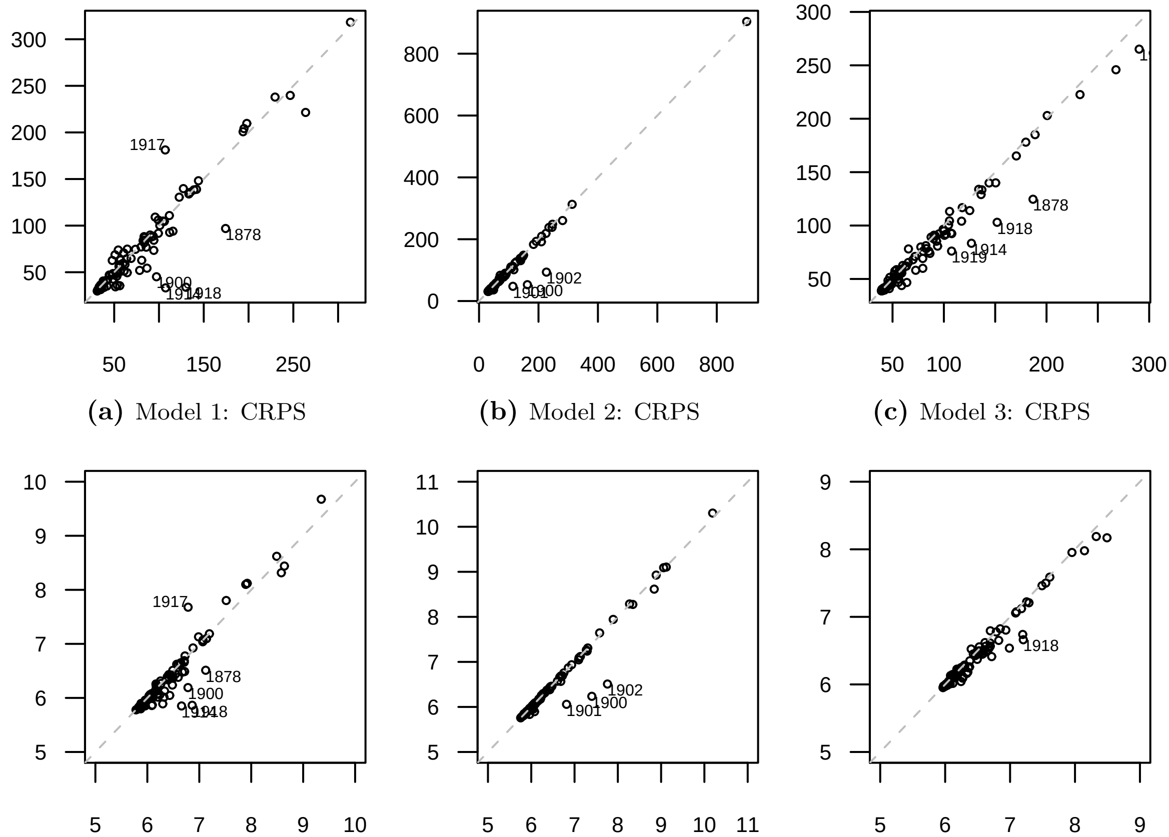
that the paired permutation test clearly rejects the hypotheses of same location parameters in the exact and approximate CRPS values on a 5% significance level (p-values $2.67 \cdot 10^{-2}$ for model 1, $2.9 \cdot 10^{-3}$ for model 2 and $1 \cdot 10^{-4}$ for model 3). Also if directly compare the exact and approximate log-scores of each model, the formal test shows the conservativeness of the approximate log-scores. So the impression from Figure 3.33 was apparently slightly misleading. However, the ranking of the models is unchanged in the approximate mean scores: both in the exact and the approximate results, the CRPS ranks model 1 best, followed by model 3 and model 2. The exact log-score ranks model 2 almost equal to model 3: Since the mean one-step-ahead log-score is equivalent to the marginal likelihood, we see from the values given on page 3.6.2 that model 2 is ranked slightly better by the exact log-score. The approximate log-score slightly favours model 3, but model 1 is still ranked highest. Therefore, the model choice using one of these two scoring rules for one-step-ahead predictive assessment would not be changed when the lightweight sampling scheme is used.

### Leave-one-out predictive assessment

We will examine the performance of the approximate leave-one-out strategy for this example of a Normal-Normal-Gamma change point model.

First, we generate 10 000 parameters samples, both from the exact and the approximate leave-one-out distributions, for all three models. Altogether, this takes 149, 119, 275 seconds for the exact sampling and 44, 16, 163 seconds for the approximate sampling, for the three different models, respectively.

***Figure 3.33*** – *Comparison of exact (x-axis) and approximate (y-axis) scores for one-step-ahead prediction in the three change point models (columns). The panels in the upper row compare the CRPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact and approximate score values exceeds 25 (CRPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: CRPS

**(b)** Model 2: CRPS

**(c)** Model 3: CRPS

**(d)** Model 1: log-score

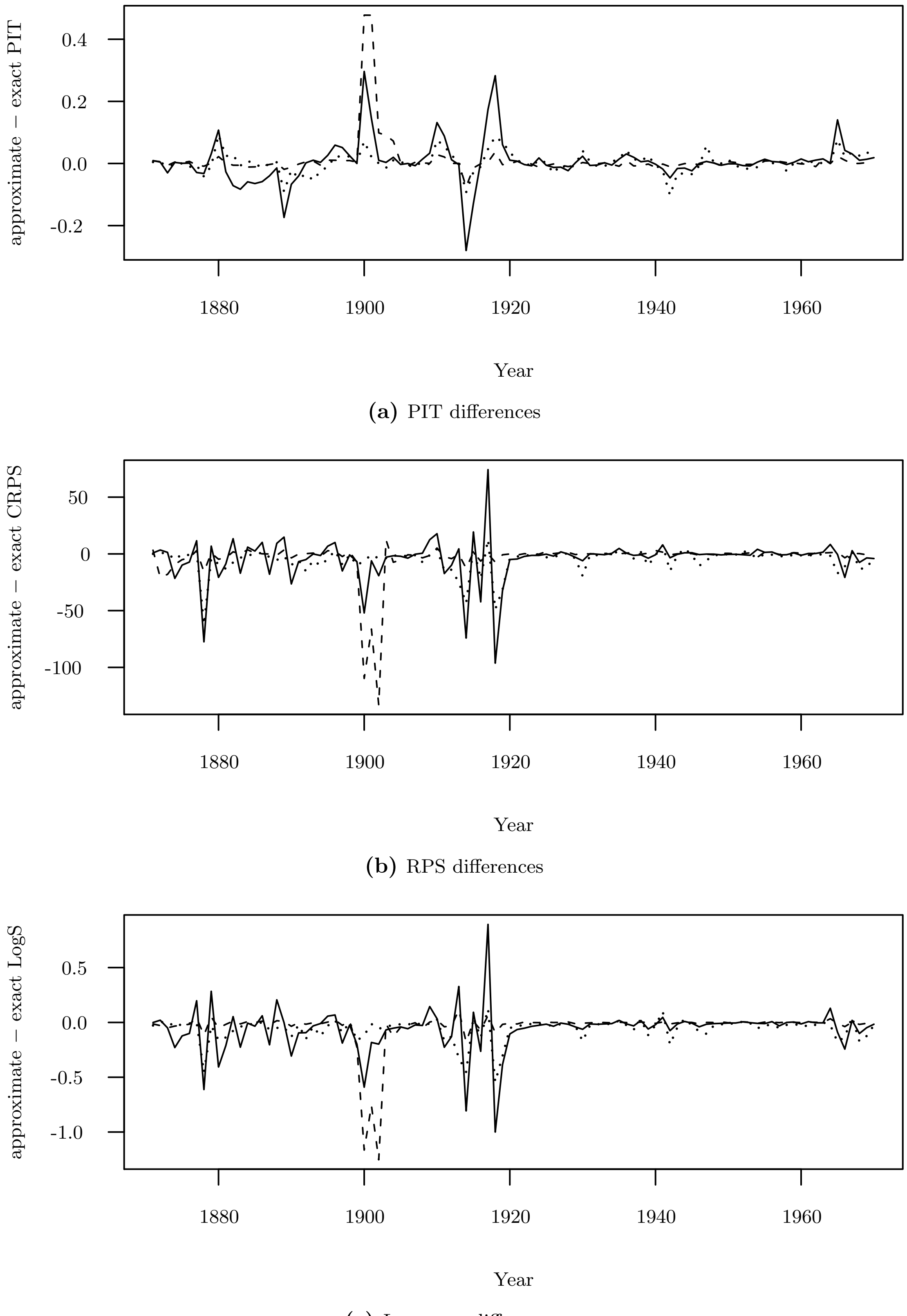
**(e)** Model 2: log-score

**(f)** Model 3: log-score

***Table 3.5*** – *Mean continuous ranked probability and logarithmic scores for the one-step-ahead prediction of the three models, under the exact and approximate sampling schemes.*

| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| CRPS | exact | 80.85 | 94.33 | 82.96 |
| | approximate | 76.65 | 90.25 | 77.66 |
| log-score | exact | 6.41 | 6.47 | 6.47 |
| | approximate | 6.35 | 6.42 | 6.41 |

***Figure 3.34*** – *Differences of the approximate and exact PIT values, continuous ranked probability and logarithmic scores for the one-step-ahead prediction, for model 1 (____), model 2 (_ _ _) and model 3 (........).*



**(a)** PIT differences

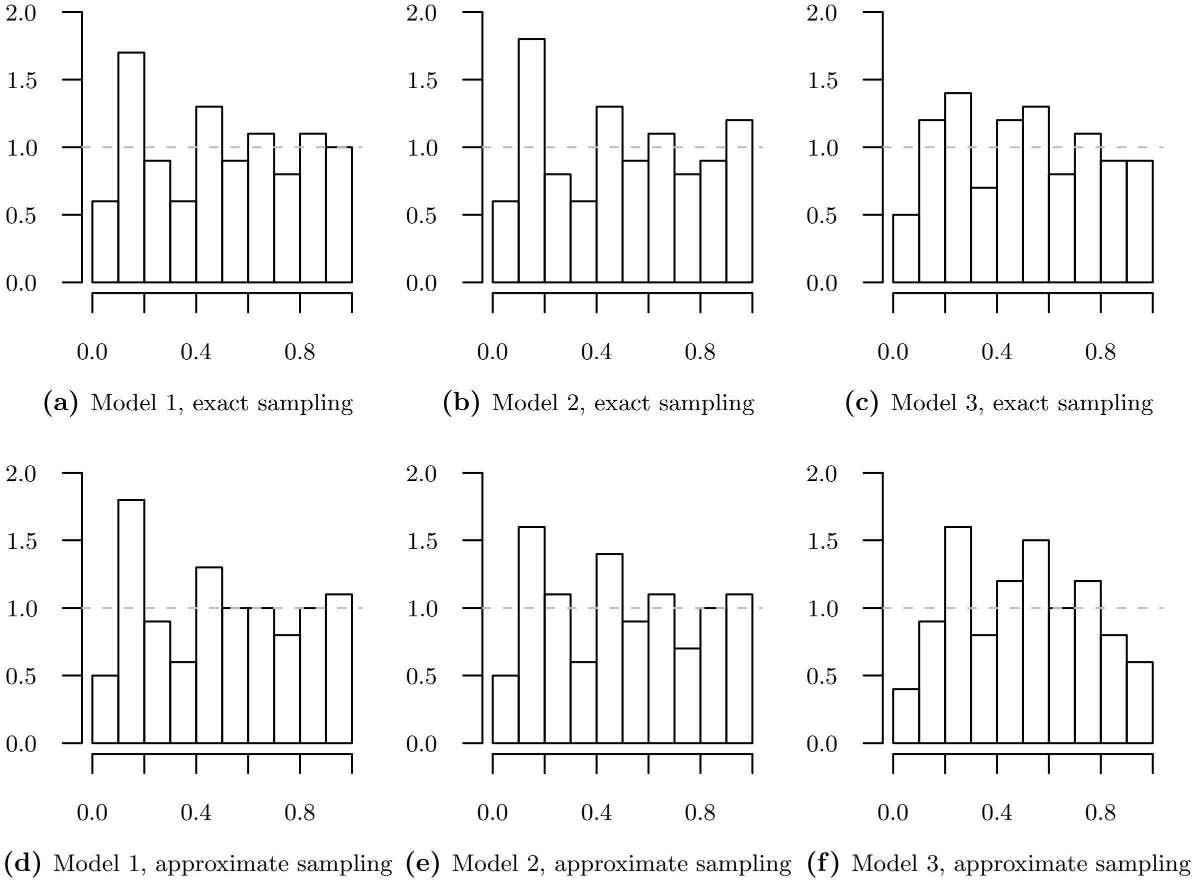
**(b)** RPS differences

**(c)** Log-score differences

Second, for each parameter sample $\boldsymbol{\xi}_t$, we generate a normal variate from the corresponding Gaussian distribution. It is a sample from the (approximated) leave-one-out predictive distribution $F_t$ for time $t$ given all other times.

The PIT histograms are presented in Figure 3.35, and do not show perfectly calibrated forecasters. Similarly to the one-step-ahead assessment, model 3 in panel (c) shows a tendency towards overdispersion. This impression is preserved by the approximate histogram in panel (f). For model 1 in panel (a) and model 2 in panel (b), the histograms could be described as left-skewed with the second bins $[0.1, 0.2]$ as outliers. The approximate results in panels (d) and (e) share this characteristic.

***Figure 3.35*** – *PIT histograms for calibration assessment of the leave-one-out prediction in the three change point models (columns). The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*

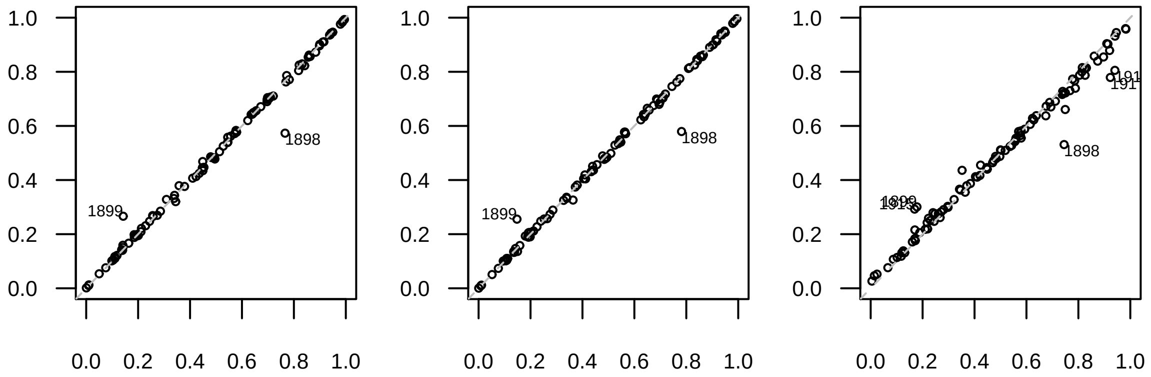


**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

The PIT values from the exact and approximate sampling schemes are compared in Figure 3.36. The approximations work very well, for all models. Only two greater deviations are visible for model 1 in panel (a) and for model 2 in panel (b). For model 3 in panel (c), few differences greater than the labelling threshold 0.1 are observed.

***Figure 3.36*** – *Comparison of exact (x-axis) and approximate (y-axis) PIT values for calibration assessment of the leave-one-out prediction in the three change point models. At most 5 time points where the absolute difference between the two values exceeds 0.1 are labelled.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

The exact and approximate scores of both proper scoring rules are compared in Figure 3.37. The approximations are very good for model 1 and model 2 scores: only the two years 1898 and 1899 before the new MAP model block are underestimated in panels (a), (d) and (b), (e), while the scores for the other years match the exact scores well. The picture is different for model 3 in panels (c) and (f). Here especially years with large exact scores (meaning bad prediction of the corresponding discharge values) yield too low approximate scores.

The mean scores for the proper scoring rules assessment of the leave-one-out prediction are summarized in Table 3.6 on page 82. The underestimation of large score values in model 3 leads to underestimated mean scores for this model. Therefore, the approximate approach ranks model 3 best for the leave-one-out prediction, while the exact sampling ranks model 3 worst and favours the other models.

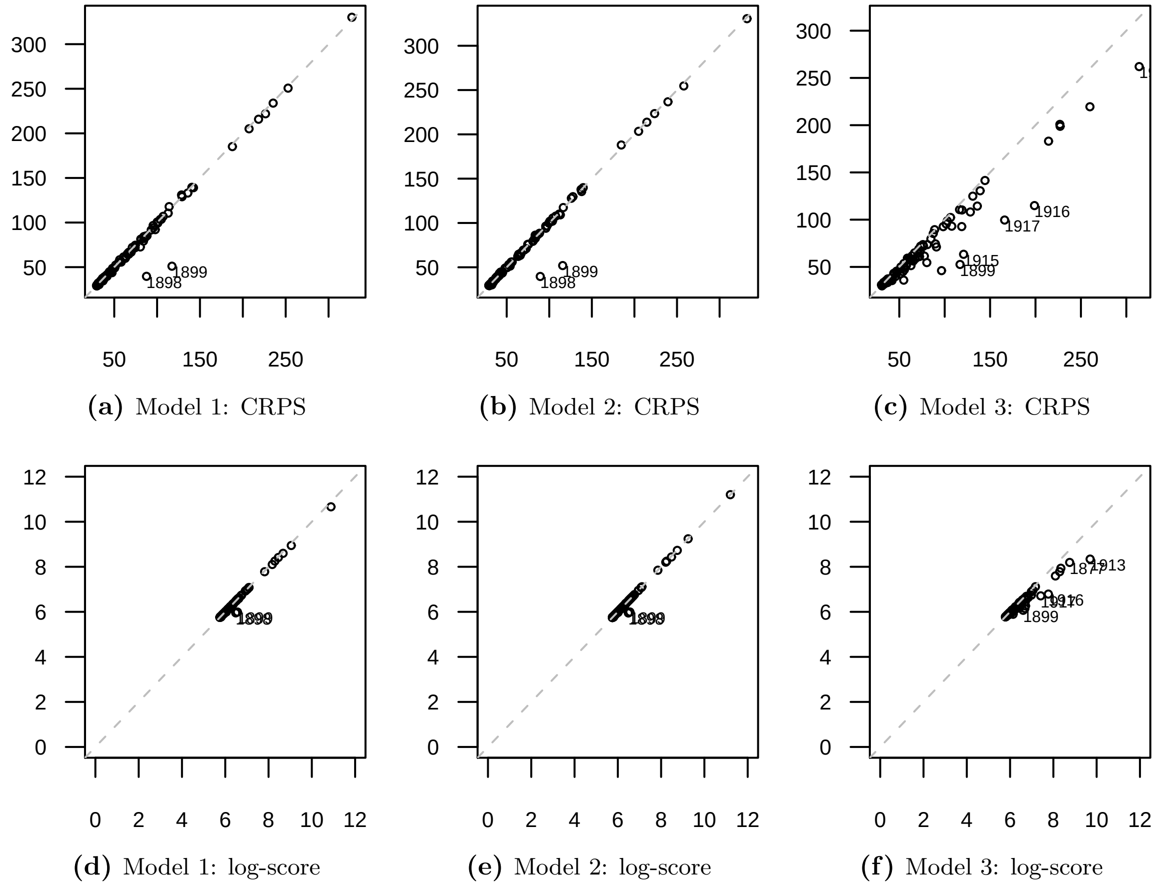
### Posterior-predictive checking

For comparison, we will look at the results of posterior-predictive model checking.

The PIT histograms are shown in Figure 3.38. While for model 1 and model 2, panels (a) and (b) pretty much agree with the exact and approximate leave-one-out PIT histograms from Figure 3.35, model 3 in panel (c) is being diagnosed a severe overdispersion by the posterior-predictive approach. This is in accordance with the closer fit to the given data.

If we compare the individual PIT values between the exact leave-one-out and the

***Figure 3.37*** – *Comparison of exact (x-axis) and approximate (y-axis) scores for leave-one-out prediction in the three change point models (columns). The panels in the upper row compare the CRPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact and approximate score values exceeds 25 (CRPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: CRPS

**(b)** Model 2: CRPS

**(c)** Model 3: CRPS

**(d)** Model 1: log-score

**(e)** Model 2: log-score

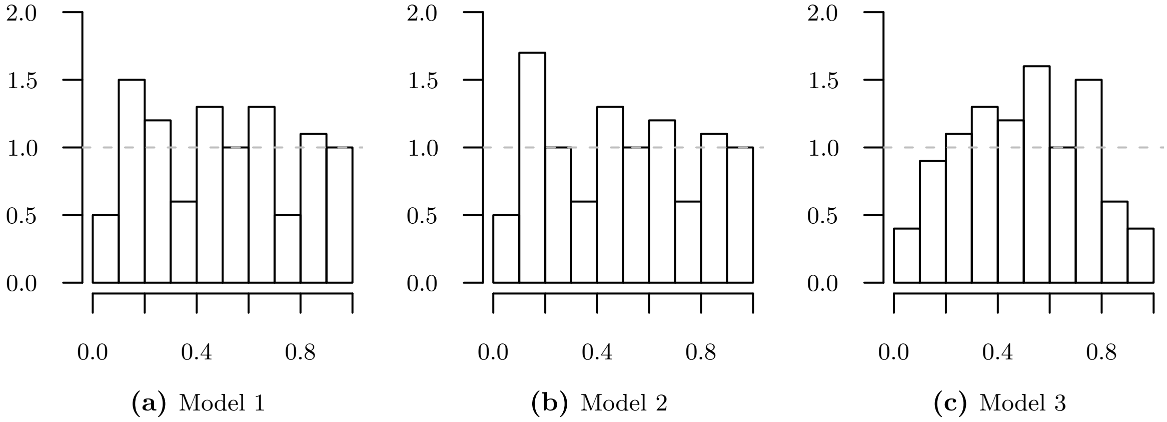
**(f)** Model 3: log-score

posterior-predictive sampling schemes in Figure 3.39, substantial shrinkage of the PIT values towards 0.5 can be seen for the model 3 PIT values in panel (c). For model 1 in panel (a) and model 2 in panel (b), the approximation by the posterior-predictive PIT values is surprisingly good.

The exact leave-one-out scores are compared with the posterior-predictive scores in Figure 3.40. For model 1 and model 2 in panels (a), (d) and (b), (e), we see that the posterior-predictive scores approximate small leave-one-out score values well. However, for large score values, the posterior-predictive scores are considerably below the exact scores. For model 3 in panels (c) and (f) the bias is already visible for small score values.

The mean scores are summarized and compared to the leave-one-out scores in Table 3.6. The heavy bias of individual model 3 posterior-predictive scores which we observed in

***Figure 3.38*** – *PIT histograms for posterior-predictive checking the three change point models.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

Figure 3.40 is mirrored in the corresponding mean CRPS and log-scores: if we only looked at the mean posterior-predictive model scores, model 3, the model with the most variable fit, appears to be much better than the other two models. Yet, using the exact and also the approximate mean scores, the difference between model 3 and model 1 is smaller.

***Table 3.6*** – *Mean continuous ranked probability and logarithmic scores for the three models, under the exact and approximate leave-one-out and the posterior predictive sampling schemes.*

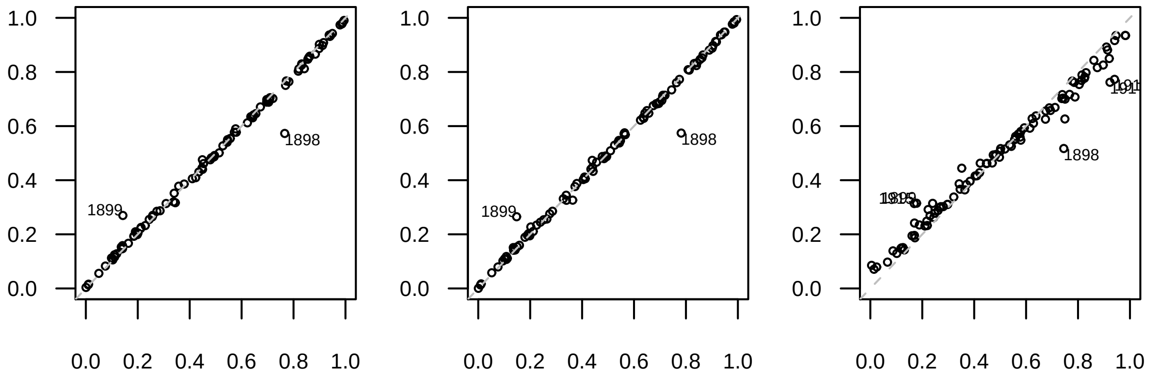
| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| CRPS | exact leave-one-out | 73.98 | 73.88 | 75.15 |
| | approximate leave-one-out | 72.02 | 72.55 | 65.57 |
| | posterior-predictive | 69.97 | 70.90 | 58.13 |
| log-score | exact leave-one-out | 6.30 | 6.31 | 6.33 |
| | approximate leave-one-out | 6.28 | 6.30 | 6.21 |
| | posterior-predictive | 6.24 | 6.26 | 6.11 |

### Results

While model 1 is clearly preferred by the marginal likelihood and the one-step-ahead predictive assessment, model 2 shows a similar performance in the leave-one-out predictive assessment. Model 3 is not preferred by any of these exact model choice criteria.

The situation is slightly different for the approximate results: Only for the one-step-ahead assessment, model 1 is still preferred, while the approximate leave-one-out scores favour model 3. However, the approximate PIT histograms for model 3 still hinted at a

***Figure 3.39*** – *Comparison of exact leave-one-out (x-axis) and posterior-predictive (y-axis) PIT values in the three change point models. At most 5 time points where the absolute difference between the two values exceeds 0.1 are labelled.*

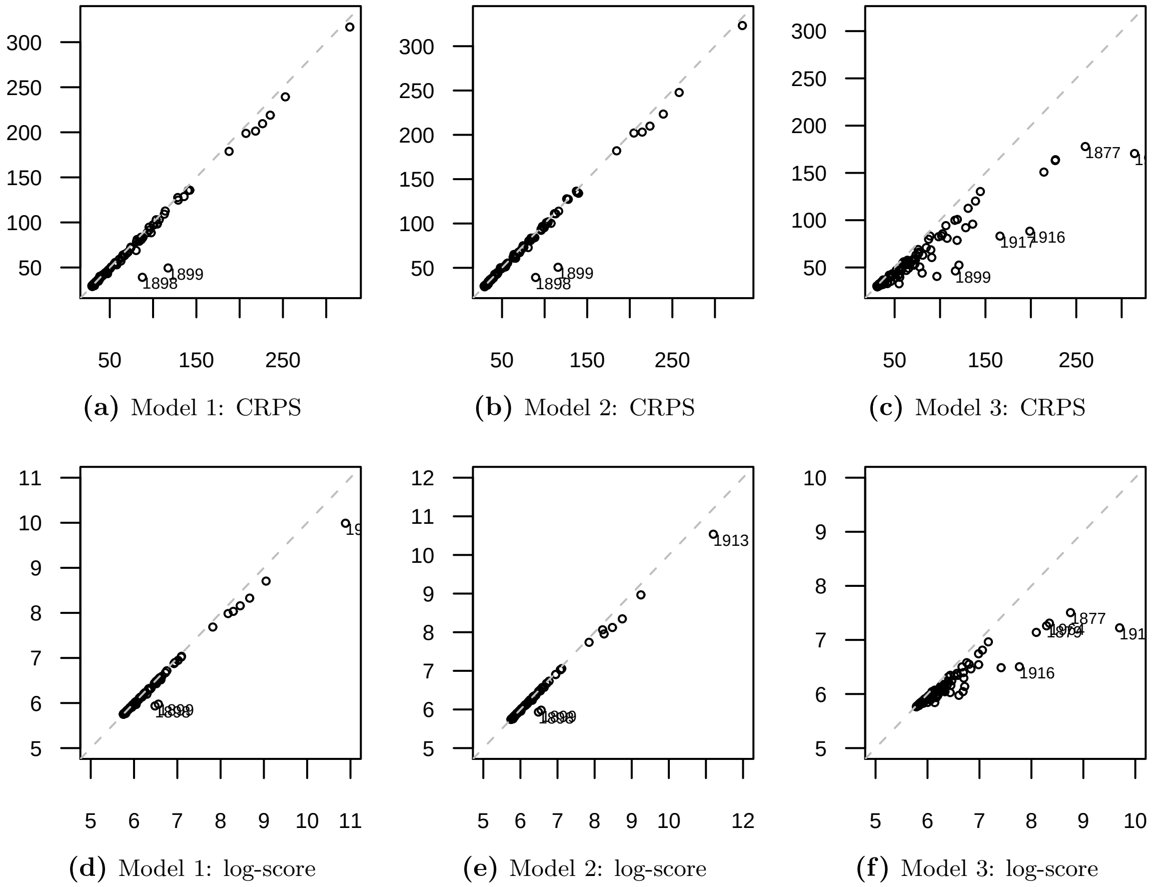


**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

possible overdispersion of model 3.

When we interpret the posterior-predictive results correctly as goodness-of-fit measures, the scores seems reasonable: model 3 has the best fit, because it follows the data-points more tightly than the other models. Also the posterior-predictive PIT histogram shows that only few of the p-values fall into the outer bins $[0, 0.1]$ and $[0.9, 1]$. Such p-values would signal that the materialized observations were extreme compared to the fitted posterior-predictive distribution at the respective time points. So the absence of many extreme p-values suggests a good fit of model 3, to the known data. Yet, the results must not be interpreted as approximations to the exact leave-one-out results, which measure the "goodness-of-prediction" for new data.

***Figure 3.40*** – *Comparison of exact leave-one-out (x-axis) and posterior-predictive (y-axis) scores in the three change point models (columns). The panels in the upper row compare the CRPS values, while the panels in the lower row compare the log-scores. At most 5 time points where the absolute difference between the exact leave-one-out and posterior-predictive score values exceeds 25 (CRPS) or 0.5 (log-score) are labelled.*



**(a)** Model 1: CRPS

**(b)** Model 2: CRPS

**(c)** Model 3: CRPS

**(d)** Model 1: log-score

**(e)** Model 2: log-score

**(f)** Model 3: log-score

## 3.7 Genetic data application

We will analyze the GC composition data introduced by Fearnhead and Vasileiou (2009, p. 133), which comprises the proportion of DNA bases that are Guanine (G) or Cytosine (C) as opposed to Adenine or Thymine in 3 kb windows of the human chromosome 1 from position 6 Mb to position 12 Mb. We computed the data from Build 35 of the finished human genome assembly (hg17, May 2004) by the International Human Genome Project for chromosome 1[i], and show the time series of length $n = 2000$ in Figure 3.41.

We compare our models with the "IsoFinder model", which is defined through the change points inferred by the IsoFinder program (Oliver, Carpena, Hackenberg, and Bernaola-Galván 2004). Its precomputed results have been obtained from the Internet[ii]. As IsoFinder reports single bases as change points, we have to round the values. For example, if IsoFinder defines 6 355 231 as a change point base (meaning that starting from base 6 355 232 a new isochore begins), we convert it to the change point index 118, because this means that we start a new block from the 119th data point, which has been aggregated from bases 6 354 001 – 6 357 00. We arrive at 115 change points, which are included in Figure 3.41.

### 3.7.1 Model fitting

Unlike Fearnhead and Vasileiou (2009, p. 135), who used (finite mixtures of) normal distributions for modelling the GC content, we could use the binomial model with $n_t = 3000$ samples and $y_t$ being the number of bases G or C in window $t = 1, \dots, n = 2000$. This would have several advantages: First, this likelihood is better fitted to the data generating process, and we can be sure that predicted GC percentages will always lie in the interval $(0, 1)$. Second, we would need to specify fewer hyperparameters for the beta prior than for the normal-gamma prior. However, the large sample sizes $n_t$ turned out as being problematic for this segmentation task, because the data was interpolated by the probabilities trend. This is due to the high information contained in the data points, which overwhelms even binomial change points priors with very small parameter $\pi$. Another possibility would be to go back to the original DNA sequence and analyze the corresponding binary time series. Yet, this is unfeasible because of the sheer length of 6 000 000 bases (6 Mb).

Therefore we stick to the normal approximation used by Fearnhead and Vasileiou (2009). We will compare models with prior settings similar to those in section 3.6.2, where we fix the hyperparameter $\nu = 0.487$ at the marginal mean of the time series. Moreover, we

[i] `http://hgdownload.cse.ucsc.edu/goldenPath/hg17/chromosomes/chr1.fa.gz`

[ii] `http://bioinfo2.ugr.es/isochores/GB/hg17/iso_chr1.html`

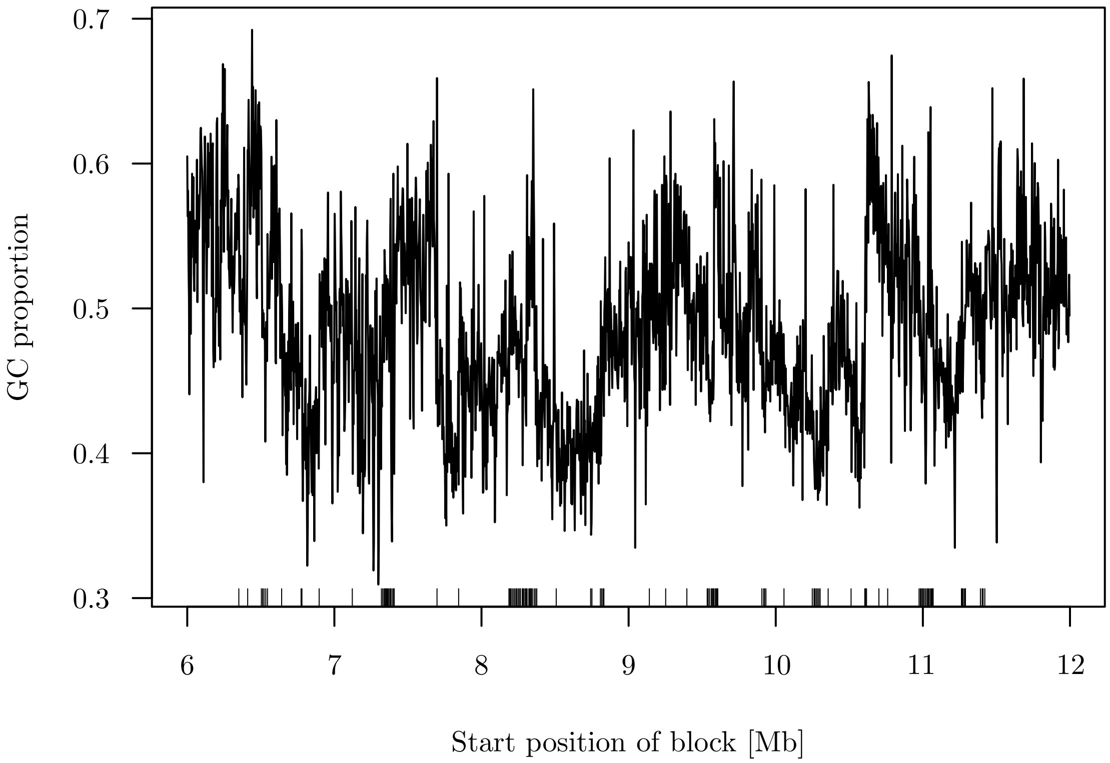


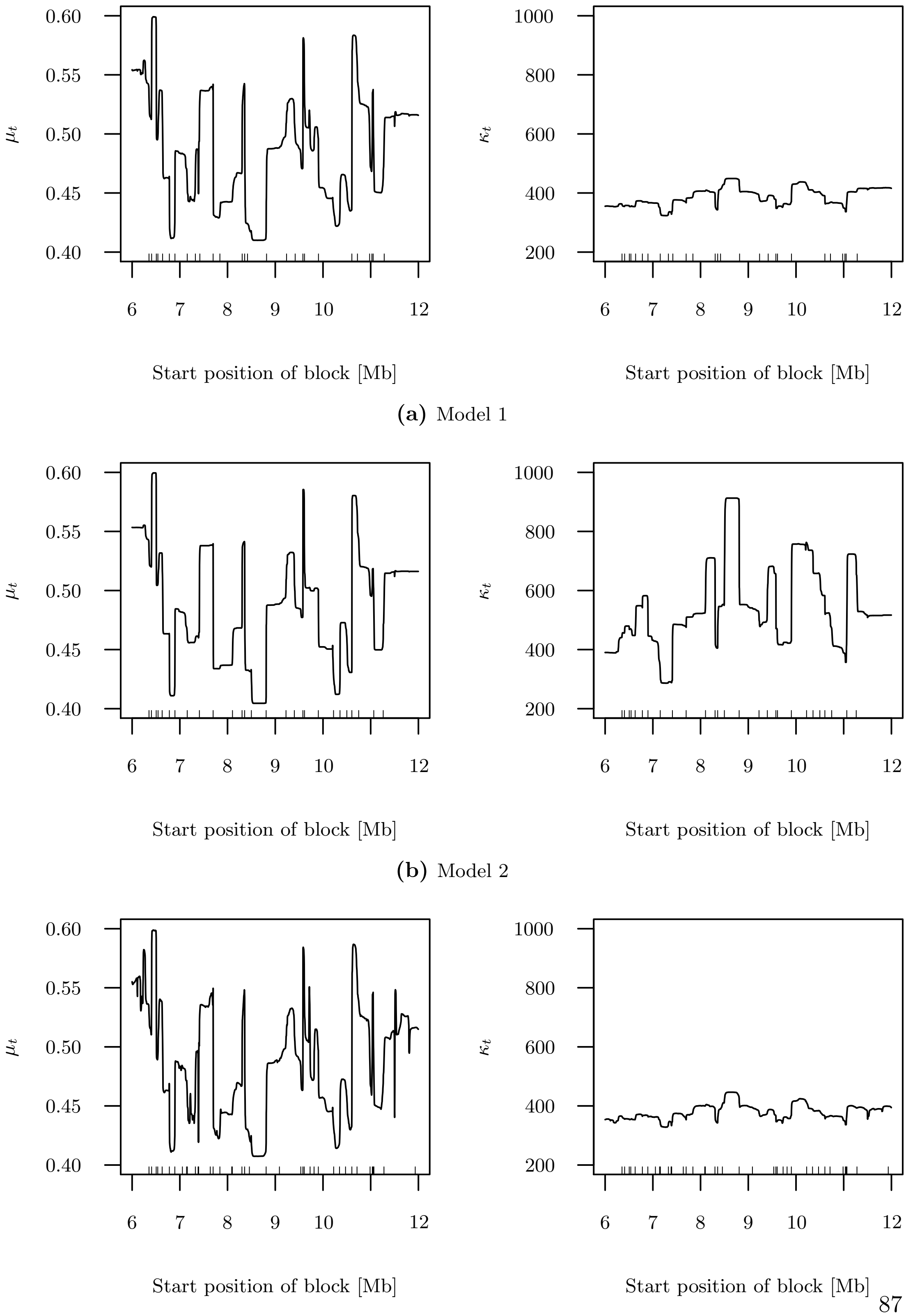
***Figure 3.41*** – *GC composition data: Proportion of G and C bases in 3 kb windows. The ticks at the bottom symbolize the change points of the IsoFinder model.*

keep the expectation of the variances at 0.003 throughout the different models, which is slightly lower than the marginal variance 0.0039.

The first model we will fit to the data uses the flat number prior for the change points, and hyperparameters $\lambda = 0.1, \alpha = 100 + 1, \beta = 0.003 \cdot 100$ for the parameters prior. The second model we want to assess also uses the flat number prior for the change points, but with hyperparameters $\lambda = 0.001, \alpha = 10 + 1, \beta = 0.003 \cdot 10$ for the parameters prior. The last model we consider uses the binomial number prior with probability $\pi = 0.05$ for a change point between any two GC windows of the time series, and the same parameters prior hyperparameters as for the first model.

We have produced 5000 samples each from the posterior distributions. Probably due to the great length of the time series, compared to the previous smaller examples, it was necessary to store the conditional densities (3.2.9) from page 23 as long double (96 bits) floating point numbers instead of double (64 bits). This was easy because we do the change points sampling in `C++`, but would have been more difficult in `R`. The memory requirements mentioned on page 26 have to be increased, but the threshold of $n = 400$ is

***Figure 3.42*** – *Posterior parameters trends for the three change point models: Both the mean (left panels) and the precision trends (right panels) are given. The change point locations in the respective MAP models are marked with ticks above the x axis. Credible intervals have been omitted for clarity.*



**(a)** Model 1

**(b)** Model 2

**(c)** Model 3

retained.

The estimated parameters trends and the change point locations in the MAP model are shown in Figure 3.42. The two models with the flat change points prior are similar: Both model 1 in panels (a) and model 2 in panel (b) have 27 MAP model change points, at similar positions in the sequence. The posterior probabilities for these MAP configurations are $2.06 \cdot 10^{-25}$ and $1.37 \cdot 10^{-17}$, respectively. The model averaged mean trend for model 1 is slightly more variable than the model 2 mean trend. Model 3 with the binomial change points prior in panel (c) exhibits a more variable mean trend. The MAP model here has probability $2.9 \cdot 10^{-48}$ and contains a total of 40 change points.

In Figure 3.43, the pointwise change point probabilities of the three models are compared to the IsoFinder change points. Overall, there is visible agreement between the two algorithms. Yet, the posterior probabilities are much more informative than the IsoFinder result, which produces more change points for regions where the conjugate change point model gives only few positions high probabilities for change points (e. g. the block between 8 Mb and 9 Mb).

The log marginal likelihood values $\log f(\boldsymbol{y})$ of the three change point models are 3256.799, 3272.511 and 3240.577, respectively. So if we should decide on the basis of the marginal likelihood, model 2 would be our best choice. However, we want to do a predictive model assessment, and are especially interested in the leave-one-out predictive assessment. Unfortunately, the exact procedure would be practically infeasible, because the change points sampling alone took 2973, 2726 and 2478 seconds for the model fitting of the three proposals, respectively. If we did an exact leave-one-out assessment, we would therefore have to wait 69, 63 and 57 days for the result!

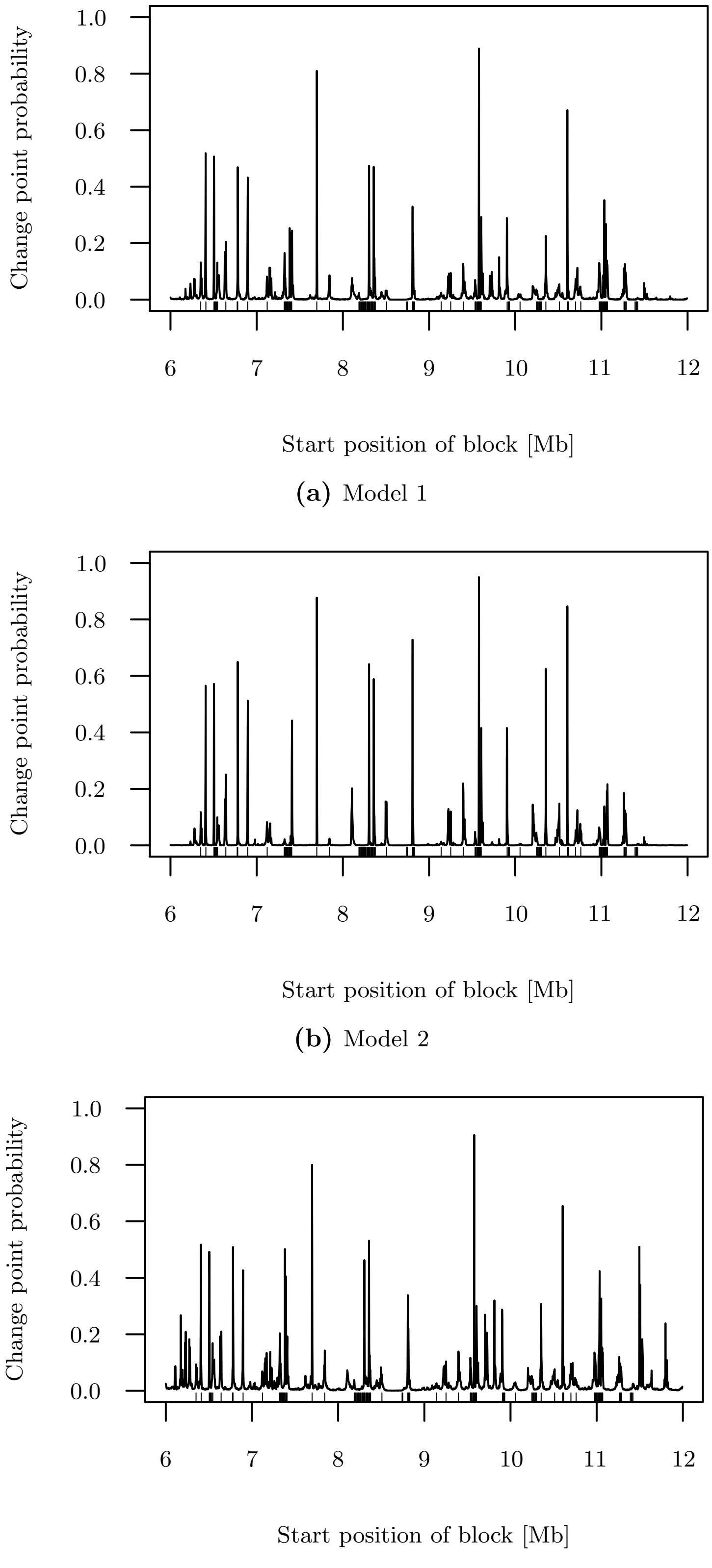
### 3.7.2 Leave-one-out predictive assessment

Thus, we will do only the approximate leave-one-out assessment for this real-world application.

Sampling from the approximate leave-one-out parameters distributions takes 4591, 4442 and 4386 seconds for the three different models, respectively. Unfortunately, numerical difficulties occurred for model 1, where some sampled mean parameters $\mu_t$ were abnormally large, even resulting in some missing values. Therefore, we set all `NA`s and mean values lower than the 0.0001 or the 0.9999 quantile to the mean of the other samples for the same time. However, this had only to be done for 1596 out of 10 000 000 values, and should thus have no relevant effect on the results. Afterwards, Gaussian random variables are produced to obtain samples from the approximate leave-one-out predictive distributions.

The resulting PIT histograms are shown in Figure 3.44. All three models show obvious overdispersion in the approximate leave-one-out predictive distributions. The degree of

***Figure 3.43*** – *Comparison of the change point probabilities in the three models and the IsoFinder change point locations (ticks at the x-axis).*



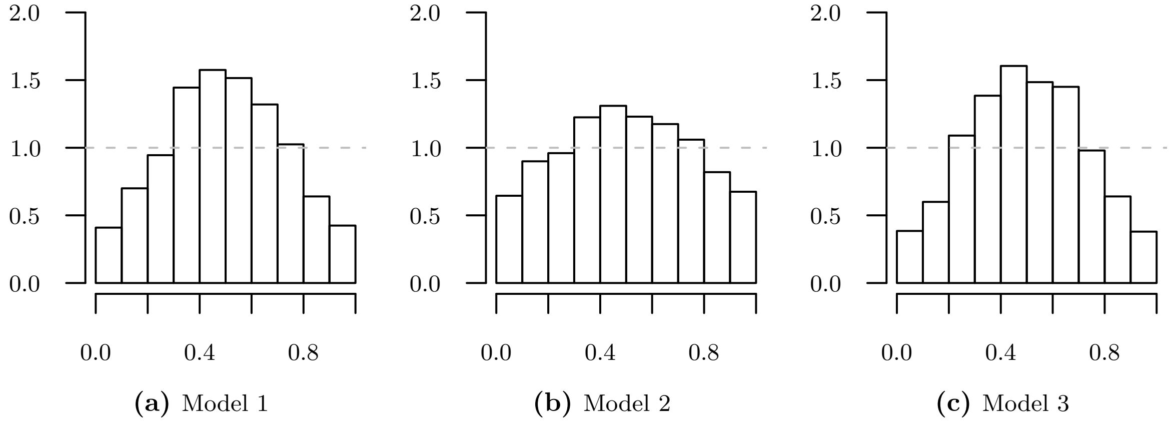
(a) Model 1

(b) Model 2

(c) Model 3

overdispersion is greater for model 3 in panel (c) than for model 1 in panel (a), and greater for model 1 than for model 2 in panel (b). So model 2 seems to have the best calibration among the three models.

***Figure 3.44*** – *PIT histograms for approximate calibration assessment of the leave-one-out prediction in the three change point models.*



(a) Model 1

(b) Model 2

(c) Model 3

The mean scores for the proper scoring rules assessment of the approximate leave-one-out prediction are summarized in Table 3.7 on page 90. While the CRPS is lowest for model 3, the log-score points to model 2 as the as the best leave-one-out predicting model.

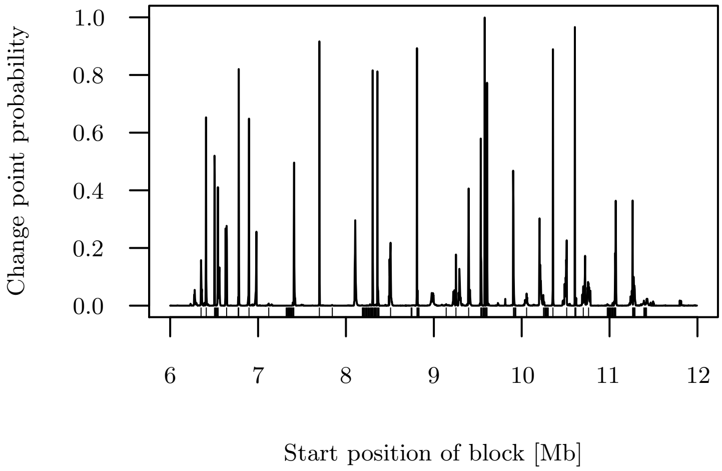
As the three examined models are all badly calibrated, we tried to change the prior parameters towards are better calibrated model. The PIT histograms in Figure 3.44 give valuable hints at what might be wrong with our current prior choice: they all show overdispersed forecast distributions. Model 2 shows a lower degree of overdispersion. If we combine this with the precision trends in Figure 3.42, where model 2 reaches levels above 800, while both other models stay below 500, we get the impression that the assumed variance mean value 0.003 is too high. As model 2 shows the most parsimonious mean trend, we might also want to specify an even sparser change points prior. The new prior choice is then $\nu = 0.487$, $\lambda = 1 \cdot 10^{-4}$, $\alpha = 11$, $\beta = 0.001 \cdot 10$ for the parameters prior and $\pi = 0.008$ for the binomial change points prior.

***Table 3.7*** – *Mean continuous ranked probability and logarithmic scores for the approximate leave-one-out prediction of the four models which have been fitted to the GC composition data.*

| Scoring Rule | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---:|---:|---:|---:|
| CRPS | 0.0235 | 0.0226 | 0.0221 | 0.0223 |
| log-score | $-1.7016$ | $-1.7824$ | $-1.7489$ | $-1.8108$ |

The MAP change points configuration in the new model has probability $6.11 \cdot 10^{-13}$ and contains a total of 28 change points. The resulting estimated parameters trends and the change point locations in the MAP model are shown in Figure 3.46. While the mean trend is very similar to that from the model 2, the precision trend exhibits a greater variability and reaches now precisions over 1400. For comparison with Figure 3.43, the change point probabilities in the new model are shown in Figure 3.45. The picture is different to the model 2 panel (b) of Figure 3.43, but the model is more sure of the change points locations: the change points probabilities are higher for some and lower for other locations. This is especially helpful at 7.4 and 11 Mb, where IsoFinder and the three previous models were very unclear about the best change points.

***Figure 3.45*** – *Comparison of the change point probabilities in the new model and the IsoFinder change point locations (ticks at the x-axis).*



The approximate PIT histogram is shown in Figure 3.47 shows the benefits of the higher precisions: the calibration is better than for model 2. The log-marginal likelihood is 3281.996, the mean approximate leave-one-out CRPS is 0.0223 and the mean log-score is −1.8108. So all model assessment tools rank the new model highest, except the CRPS which gives model 3 a minimally better score (cf. Table 3.7).

## 3.7.3 Results

The marginal likelihood, which is equivalent to the mean one-step-ahead log-score, ranks the new model 4 highest, followed by models 2, 1 and 3. The approximate leave-one-out log-score gives a very similar ranking, only the positions of the worst two models are exchanged. However, the approximate leave-one-out CRPS ranks model 3 best, followed by models 4, 2 and 1.

One possibility to somehow check the accuracy of the approximations for this concrete

***Figure 3.46*** – *Posterior parameters trends for the new change point model.*

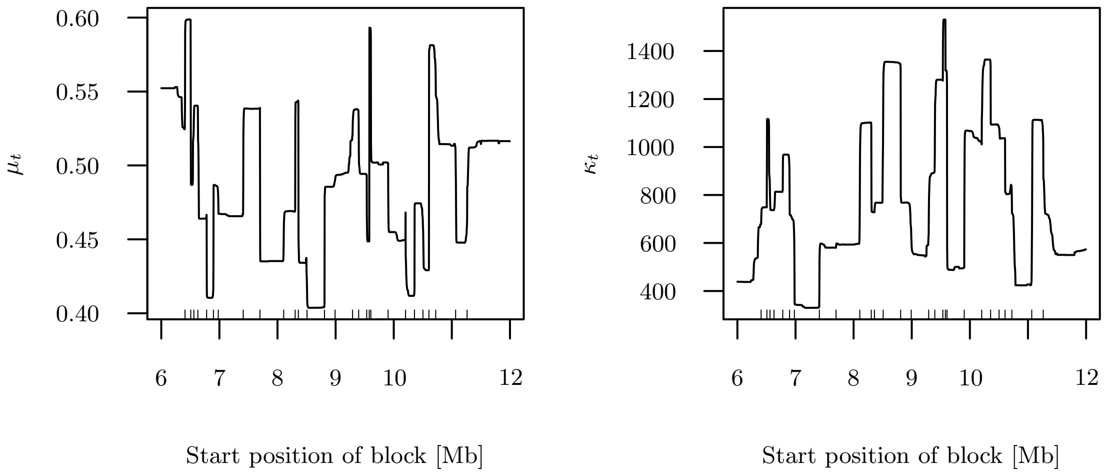
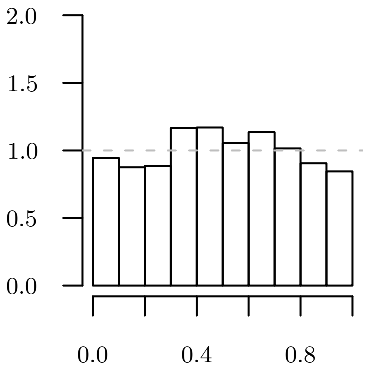


***Figure 3.47*** – *PIT histogram for approximate calibration assessment of the leave-one-out prediction in the new change point model.*



large data example would be to select a small window of 200 observations, say. For this subset, the exact results could be computed and compared with the approximate scores, similarly as we have done it in the three case studies in this chapter. Yet, even without doing this subset accuracy check, it is worth considering the ranking of the marginal likelihood values: Because they also rank the new model best, the similar result of the approximate leave-one-out log-scores is supported.

This real-world example is instructive, because the approximate leave-one-out PIT histograms could guide us from the three badly calibrated models to a better calibrated new model. This is a very important aspect of the PIT histogram, because it provides a model criticism tool – we do not only see that some model is ranked higher than the other model (using the proper scoring rules), but we do also get information about what may be wrong with the bad models, and what could remedy the deficiencies.

## 3.8 Summary

In this chapter we have investigated the performance of Marshall-Spiegelhalter type approximations to the exact one-step-ahead and leave-one-out assessment results for conjugate change point models. For three different distribution families, the case studies showed that these approximations are good and can be obtained with less computational effort, and should be used instead of the posterior-predictive results when the out-of-sample predictions of the models matter. While for the small problems in the case studies the exact computations were still feasible, the approximations were vital for predictive model assessment in the large genetic data example. For the conjugate change point models, the marginal likelihood can be computed, which also allows the computation of Bayes factors between competing models. Model choice for more general change point models could be based on the one-step-ahead and leave-one-out assessments alone, when the marginal likelihood cannot be estimated reliably. Nevertheless, Chib (1998) proposes the MCMC estimation of marginal likelihoods for comparing models with fixed numbers of change points. Fearnhead and Vasileiou (2009) avoid the MCMC convergence issues and can calculate the marginal likelihood exactly. The work in this thesis is a first step towards a more thorough predictive assessment of change point models, which is asked for by Held, Hofmann, Höhle, and Schmid (2006, p. 435) for a more complex infectious disease counts model.

Comparing the exact one-step-ahead with the exact leave-one-out model scores from the case studies, we notice that both assessment types yield the same ranking for the Poisson-Gamma example (section 3.4.2) but slightly different rankings for the binomial-beta example (section 3.5.2): while the one-step-ahead scores favour model 1 over model 3, the order is reversed by the leave-one-out scores. For the normal normal-gamma example (section 3.6.2) the CRPS and log-score rankings are equal for the one-step-ahead assessment (models 1, 3, 2), but the exact leave-one-out rankings differ (2, 1, 3 and 1, 2, 3, respectively).

The mean scores tables exhibit a common pattern, across all three examined distribution families. First, within each table (see e. g. Table 3.2 on page 49), the exact scores are always highest, followed by the approximate scores and then (for the leave-one-out prediction) the posterior-predictive scores. So the approximate scores are a bit too optimistic for the examined models, but the posterior-predictive scores are no good substitutes for the exact scores at all. Second, if we compare the scores between the one-step-ahead and the leave-one-out tables (compare e. g. Table 3.1 on page 46 with Table 3.2), we notice that the one-step-ahead scores are always higher than the respective leave-one-out scores. This means that the one-step-ahead prediction is more difficult than the leave-one-out prediction, whether it is exact or approximate. Both findings can be explained by Figure 3.4. First, the posterior-predictive sampling scheme uses the full data set, while the approximate leave-one-out scheme only partially uses the information for the predicted time, and the exact leave-one-out scheme only uses the information from all other times. Analogously for the one-step-ahead prediction, the exact approach must do without the partial ahead information used by the approximate scheme. The more data is available, the easier is the prediction, which corresponds to lower mean scores. Second, the exact one-step-ahead prediction does not use the data after the next time, unlike the exact leave-one-out prediction (except for the prediction of the last time, when both tasks coincide). For the approximate versions, the later times are used partially by the one-step-ahead but fully by the exact scheme. Therefore here also more data is available to the leave-one-out predictions, making it easier and thus producing lower mean scores than the one-step-ahead predictions.

In the PIT comparison plots, we have recognized a shrinkage of the posterior-predictive (mid-)PIT values towards 0.5 relative to the exact leave-one-out PIT values. This can easily be explained by the conservativeness of the posterior-predictive results: The corresponding predictive distributions are shrunk towards to the observation which was not known to the leave-one-out predictive distribution. Thus, the observation is less extreme relative to the forecaster, and the PIT value is shrunk towards 0.5. The shrinkage is much weaker for the approximate PIT values, because the information from the observation is only partially used to sample the change points. For example in the approximate comparison plot in panel (c) of Figure 3.24 (p. 64) the characteristic S-form can be noted. But the S-form is much clearer in the corresponding posterior-predictive comparison plot in panel (c) of Figure 3.26 (p. 66).

The PIT shrinkage explained above was strongest for overfitting models. This is natural, because the posterior-predictive distribution is more different from the leave-one-out forecast when the model adapts more strongly to the known data. Interestingly, the comparison of the posterior-predictive with the corresponding leave-one-out distribution was proposed for assessing the influence (or "leverage" in classic regression) of the individual on its own fit by the model (Gilks, Richardson, and Spiegelhalter 1998, p. 151). This question is closely related to the influence measures typically used in linear regression, for example the Cook's distance (Cook and Weisberg 1980). Furthermore, the case studies have exemplified that the models with the best fit are not necessarily the models with the best predictive performance. For example, the normal normal-gamma model 3 in section 3.6.2 fits the given data best (based on the posterior-predictive scores), but has the worst leave-one-out predictions (based on the exact leave-one-out scores).

# 4 Random effects models for longitudinal data

In section 4.1, longitudinal data models are motivated. The random effects modelling framework is detailed and specialized to the linear mixed model in section 4.2. Similarly in section 4.3, the predictive assessment schemes are first presented for general random effects models before being implemented for the linear mixed model. Section 4.4 compares the assessment results of the correct model and three wrong models in a small-scale simulation study. Real data are analyzed in sections 4.5 and 4.6: First the performances of the approximate assessment scheme are evaluated for a subset, before being applied to the full data sets. Section 4.7 summarizes the results of this chapter.

## 4.1 Introduction

In its most general definition, longitudinal data is a collection of multiple time series. By contrast, in chapter 3 we examined a model class suited to the analysis of single time series. Typically and also in our real data examples in this chapter, each time series is produced by measuring repeatedly the outcome for a single individual. In parallel, other variables are recorded which could be associated with the outcome trajectories. Longitudinal data models allow a statistical analysis which accounts for the correlation within the time series, which could be an age effect, for example. Moreover, cohort effects can be estimated, which could be responsible for different baseline levels of the time series. If the individuals enter the study over a long time, also calendar-time effects could be of interest, for example if health care has improved and is relevant for the outcome.

A book-length overview of different methods in longitudinal data analysis is given by Diggle, Heagerty, Liang, and Zeger (2002). They also present a CD4 data example from the same data pool as our CD4 example in section 4.5. In this thesis we concentrate on random effects models for longitudinal data: these account for unobserved heterogeneity between individuals by declaring the differences (the random effects) to be distributed to (typically) a normal distribution. This assumption reflects that effects which cannot be explained by observed covariates have (approximately) a normal distribution in the population. Note that the word “random” is in fact superfluous in our setting because in the proposed Bayesian inference all effects are assumed to be random, mirroring the uncertainty about them. Yet, in the frequentist inferential framework the randomness of the parameters is a striking element, so the models were coined "random effects models". In particular, the implementation will focus on the linear mixed effects models, which assume that given the population effects and random effects (thus "mixed" effects) and the covariates, a single observation has a normal distribution.

## 4.2 Modelling framework

Section 4.2.1 describes the data to which the general random effects model from section 4.2.2 can be applied. The details for the normal linear mixed effects model are given in section 4.2.3.

### 4.2.1 Data

We assume that we intend to analyze a longitudinal data set comprising $n$ individuals $i = 1, \dots, n$ with time series of scalar outcomes. For individual $i$, $n_i$ scalar outcomes $y_{ij}$ indexed by $j = 1, \dots, n_i$ are recorded. That is, the multiple time series $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})'$, $i \in \mathcal{N}$, do not need to have the same lengths. Moreover, non-equidistant measurement times $t_{ij}$ for the observations are allowed. The whole longitudinal data set is denoted as $\mathcal{Y}$. Usually covariates are recorded in parallel to the outcomes. The notation for the resulting design matrices is detailed in section 4.2.2.

### 4.2.2 Model

The general model assumes independence of the time series $\boldsymbol{y}_i$ conditional on the parameter vector $\boldsymbol{\xi}$ and the individual random effects $\boldsymbol{\alpha}_i$. So the likelihood for the observations is

$$\boldsymbol{y}_i \overset{ind}{\sim} f(\boldsymbol{y}_i \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i), \quad i \in \mathcal{N}.$$

Covariates (especially times $t_{ij}$) may also enter the data generating distribution, but are suppressed in the notation for clarity. Then the independence is understood conditional on the covariates, too.

The distribution of the random effects $\boldsymbol{\alpha}_i$ is parametrized by $\boldsymbol{\delta}$:

$$\boldsymbol{\alpha}_i \overset{iid}{\sim} f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}), \quad i \in \mathcal{N}.$$

Usually, this distribution will be a (multivariate) normal distribution. The prior for the non-individual parameters in $\boldsymbol{\xi}$ has a hyperparameter $\boldsymbol{\tau}$:

$$\boldsymbol{\xi} \sim f(\boldsymbol{\xi} \,|\, \boldsymbol{\tau}).$$

Finally, a joint hyperprior can be included for the prior parameters:

$$\boldsymbol{\delta}, \boldsymbol{\tau} \sim f(\boldsymbol{\delta}, \boldsymbol{\tau}).$$

The structure of the general model framework is summarized in Figure 4.1. Note that this structure also applies to cluster data, where the index $i$ then identifies cluster $i$ instead of individual $i$. Only the chosen likelihood with its covariates specializes the framework to longitudinal data.

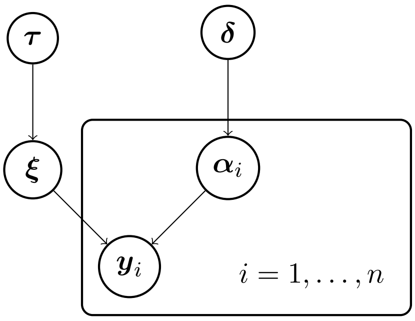


***Figure 4.1*** – *Graphical model of the proposed random effects model framework.*

### 4.2.3 Special case linear mixed model

We consider the normal linear mixed model of Laird and Ware (1982) as a special case of the random effects model from section 4.2.2.

For the real-valued vectors $\boldsymbol{y}_i \in \mathbb{R}^{n_i}$, $i \in \mathcal{N}$, the non-individual parameter is $\boldsymbol{\xi} = (\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the fixed effects vector and $\sigma^2$ is the regression variance. $\boldsymbol{\alpha}_i \in \mathbb{R}^q$ is the individual random effects vector for individual $i$. The data generating distribution for individual $i$ is then specified as

$$f(\boldsymbol{y}_i \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i) = \mathrm{N}_{n_i}(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_i, \sigma^2 \boldsymbol{I}_{n_i}),$$

where $\boldsymbol{\mu}_i := \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\alpha}_i$ is the assumed mean vector for the independent observations with common variance $\sigma^2$. $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times p}$ collects the covariate vectors $\boldsymbol{x}_{ij}$ for observation $i$, i. e. it is the design matrix

$$\boldsymbol{X}_i = (\boldsymbol{x}_{i1} \,|\, \boldsymbol{x}_{i2} \,|\, \cdots \,|\, \boldsymbol{x}_{in_i})'$$

for the fixed effects $\boldsymbol{\beta}$. Analogously, $\boldsymbol{Z}_i \in \mathbb{R}^{n_i \times q}$ is the design matrix for the random effects $\boldsymbol{\alpha}_i$.

The distribution of the random effects $\boldsymbol{\alpha}_i$ depends on a vector-valued hyperparameter, namely the variances $\boldsymbol{\delta} = (\delta_1^2, \ldots, \delta_q^2)'$ of the individual random effects distributions:

$$f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}) = \mathrm{N}_q(\boldsymbol{\alpha}_i \,|\, \boldsymbol{0}, \operatorname{diag} \boldsymbol{\delta}).$$

This assumption implies that usually the individual design matrix $\boldsymbol{Z}_i$ will be composed of some columns already present in the population design matrix $\boldsymbol{X}_i$, which are indexed by $\mathcal{Z}$, say. Then the resulting individual vector entering the $i$-th predictor $\boldsymbol{\mu}_i$ is $\boldsymbol{Z}_i\boldsymbol{\eta}_i$, where the individual effect $\boldsymbol{\eta}_i = \boldsymbol{\beta}_{\mathcal{Z}} + \boldsymbol{\alpha}_i$ sums up the population effect $\boldsymbol{\beta}_{\mathcal{Z}}$ for these covariates and the individual deviation $\boldsymbol{\alpha}_i$. In this notation we have $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}_{\bar{\mathcal{Z}}} + \boldsymbol{Z}_i\boldsymbol{\eta}_i$ where $\bar{\mathcal{Z}}$ collects the indexes of fixed effects without corresponding random effects. Of course it can be sensible to have random effects without directly corresponding population effects in the model. This is the case, for example, in model 6 in section 4.5.2.

The prior for the non-individual parameters is composed of an improper flat prior on the fixed effects $\boldsymbol{\beta}$ and an inverse-gamma distribution on the regression variance $\sigma^2$ with fixed hyperparameters $a, b \in \mathbb{R}_+$:

$$f(\boldsymbol{\xi} \,|\, \boldsymbol{\tau}) \propto \mathrm{IG}(\sigma^2 \,|\, a, b).$$

Here $\boldsymbol{\tau} = (a, b)'$ can be formally included in the model by assigning it a point-mass hyperprior at the fixed hyperparameter values. The flat prior on $\boldsymbol{\beta}$ ensures that the model can freely center these population effects at the appropriate scale, without influence of e. g. a shrinkage prior. Yet, for the random effects $\boldsymbol{\alpha}_i$ the normal shrinkage prior is necessary because otherwise the posterior would not exist: the population effects could not be distinguished from the individual effects.

The random prior parameters in $\boldsymbol{\delta}$ are assigned a non-degenerate hyperprior, namely the product of identical inverse gamma distributions with fixed hyperparameters $c, d \in \mathbb{R}_+$. So the common prior for $\boldsymbol{\delta}$ and $\boldsymbol{\tau}$ is

$$f(\boldsymbol{\delta}, \boldsymbol{\tau}) = \prod_{k=1}^{q} \mathrm{IG}(\delta_k^2 \,|\, c, d)\, \mathbb{I}_{\{(a,b)'\}}(\boldsymbol{\tau}),$$

where $\mathbb{I}_{\{(a,b)'\}}(\boldsymbol{\tau})$ denotes the density of the Dirac point measure $\delta_{(a,b)}$ in $(a, b)$ for $\boldsymbol{\tau}$.

We will focus on models of this normal linear mixed effects model type. Specifically, the open source program `BayesX`[i] implements the methodology with MCMC based posterior inference. Bayesian MCMC inference with Gibbs sampling is described in the `BayesX` methodology manual (Belitz, Brezger, Kneib, and Lang 2009a, section 6.1.1, p. 21). See the reference manual by Belitz, Brezger, Kneib, and Lang (2009b) on pp. 67 and 70 for the specification of random intercept and random slope terms, respectively. `BayesX` uses "hierarchical centring" reparametrisations which often improve convergence of the MCMC samples (Gelfand, Sahu, and Carlin 1995). Essentially this means that in the sampling scheme, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{Z}}', \boldsymbol{\beta}_{\bar{\mathcal{Z}}}')'$ and $\boldsymbol{\eta}_i$ $(i = 1, \dots, n)$ are sampled instead of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_i$. The term stems from the fact that the individual effect $\boldsymbol{\eta}_i$ is centered around the population effect

[i] `http://www.stat.uni-muenchen.de/~bayesx/bayesxdownload.html`

$\boldsymbol{\beta}_{\mathcal{Z}}$. Of course, the original random effect samples can be recovered using the identity $\boldsymbol{\alpha}_i = \boldsymbol{\eta}_i - \boldsymbol{\beta}_{\mathcal{Z}}$.

## 4.3 Exact and approximate predictive assessment

This section introduces the three different predictive assessment schemes: the exact and approximate cross-validation assessment schemes are presented in section 4.3.1. Section 4.3.2 contrasts this with the posterior-predictive assessment, which can only be used for goodness-of-fit assessment. The implementation of the three different schemes for the linear mixed model is detailed in section 4.3.3.

### 4.3.1 Cross-validation assessment

The leave-one-out cross-validation for random effects models is computationally demanding, because each individual $i$ is left out in turn. The model must then be fitted to the reduced longitudinal data set of size $n-1$ to obtain a prediction (in the form of samples from the predictive distribution) for the left out time series $\boldsymbol{y}_i$. The fitting process often and also for the linear mixed model is based on an MCMC sampling scheme, which requires much computing time. Therefore we will propose an approximate leave-one-out sampling scheme, which should produce results close to those of the exact leave-one-out sampling scheme, while easing the computational burden. Note that we leave out whole time series but not individual scalar observations in the cross-validation.

We have presented the cross-validation procedure as a leave-one-out procedure with respect to the $n$ vector-valued observations. However, one could also interpret it as an $n$-fold cross-validation of the $\sum_{i=1}^{n} n_i$ individual scalar observations, where the test sets are identical to the clusters. In this view, we can in principle apply scalar checking tools, e. g. compute PIT values for the individual observations. This could be valuable, because there is no direct generalization of the PIT for vector-valued observations. The resulting PIT histograms can then be compared to the BOT histograms.

#### Exact sampling

The exact leave-one-out predictive density, for the prediction of $\boldsymbol{y}_i$ from the remaining observations $\mathcal{Y}_{\mathcal{N}\setminus\{i\}}$, is given by

$$\begin{aligned} f(\boldsymbol{y}_i \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}}) &= \iiiint f(\boldsymbol{y}_i, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\alpha}_i \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})\, d\boldsymbol{\delta}\, d\boldsymbol{\tau}\, d\boldsymbol{\xi}\, d\boldsymbol{\alpha}_i \\ &= \iiiint f(\boldsymbol{y}_i \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i) f(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \boldsymbol{\alpha}_i \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})\, d\boldsymbol{\delta}\, d\boldsymbol{\tau}\, d\boldsymbol{\xi}\, d\boldsymbol{\alpha}_i \\ &= \iiiint f(\boldsymbol{y}_i \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i) f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}) f(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi} \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})\, d\boldsymbol{\delta}\, d\boldsymbol{\tau}\, d\boldsymbol{\xi}\, d\boldsymbol{\alpha}_i. \end{aligned}$$

The last equation follows because

$$f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi}, \mathcal{Y}_{\mathcal{N}\setminus\{i\}}) = f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}),$$

which can be read off the graphical model in Figure 4.2: Each node in the graph is conditionally independent of all non-descendant nodes, given all parent nodes. Since $\boldsymbol{\delta}$ is the only parent of $\boldsymbol{\alpha}_i$, and $\boldsymbol{\tau}$, $\boldsymbol{\xi}$ and $\mathcal{Y}_{\mathcal{N}\setminus\{i\}}$ are non-descendants, the statement follows.

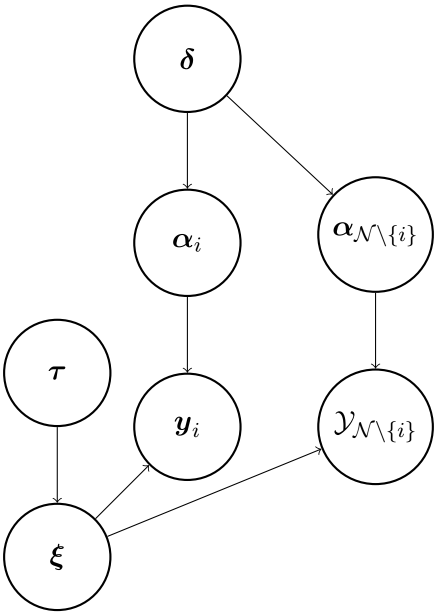


***Figure 4.2*** – *Graphical model of the leave-one-out setting.*

Thus, sampling from the exact leave-one-out predictive density $f(\boldsymbol{y}_i \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})$ proceeds as follows:

1. Draw $\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi}$ from the reduced posterior obtained from the reduced input data $\mathcal{Y}_{\mathcal{N}\setminus\{i\}}$.

2. Draw the random effect $\boldsymbol{\alpha}_i$ from the random effects distribution with parameter $\boldsymbol{\delta}$ being the sample from above.

3. Draw the prediction sample $\boldsymbol{y}_i^*$ from the data generating distribution with the sampled parameters $\boldsymbol{\xi}, \boldsymbol{\alpha}_i$.

### Approximate sampling

The exact leave-one-out cross-validation will be infeasible for normally sized data sets, when the reduced posterior sampling is computationally demanding. The approximate leave-one-out predictive density thus replaces the reduced posterior $f(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi} \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})$ with the full posterior $f(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi} \,|\, \mathcal{Y})$:

$$\tilde{f}(\boldsymbol{y}_i \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}}) := \iiiint f(\boldsymbol{y}_i \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i) f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}) f(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi} \,|\, \mathcal{Y}) \, d\boldsymbol{\delta} \, d\boldsymbol{\tau} \, d\boldsymbol{\xi} \, d\boldsymbol{\alpha}_i. \tag{4.3.1}$$

Sampling from $\tilde{f}(\boldsymbol{y}_i \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})$ proceeds as follows:

1. Draw $\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\xi}$ from the *full* posterior obtained from the *full* input data $\mathcal{Y}$.

2. Identical to the exact sampling: Draw the random effect $\boldsymbol{\alpha}_i$ from the random effects distribution with parameter $\boldsymbol{\delta}$ being the sample from above.

3. Identical to the exact sampling: Draw the prediction sample $\boldsymbol{y}_i^*$ from the data generating distribution with the sampled parameters $\boldsymbol{\xi}, \boldsymbol{\alpha}_i$.

Thus, only one MCMC run is necessary for a leave-one-out cross-validation, where in turn each observation is left out from the input data, and predicted from the remaining data: Just run a single MCMC chain, save the samples of $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ and do steps 2 and 3 for each of the samples.

Note that the notation in this chapter is deliberately close to the notation in the mixed predictive checking section in Fahrmeir and Kneib (2010, ca. p. 178). This chapter is actually a first step to "assessing the quality of full-data mixed predictive checking" asked for by the authors.

### 4.3.2 Goodness-of-fit assessment

Analogous to the argumentation in section 3.3.3 for conjugate change point models, we can use posterior-predictive samples to check the goodness-of-fit of random effects models. Obtaining these samples $\mathcal{Y}^*$ from

$$f(\mathcal{Y}^* \,|\, \mathcal{Y}) = \iint \prod_{i=1}^{n} f(\boldsymbol{y}_i^* \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i) f(\boldsymbol{\xi}, \boldsymbol{\alpha}_i \,|\, \mathcal{Y}) \, d\boldsymbol{\xi} \, d\boldsymbol{\alpha}$$

is easy: For each posterior sample $(\boldsymbol{\xi}, \boldsymbol{\alpha})$, draw the replicate $\boldsymbol{y}_i^*$ from the likelihood for all individuals $i = 1, \ldots, n$.

The fundamental difference to the approximate leave-one-out sampling scheme described in section 4.3.1 is that the random effects $\boldsymbol{\alpha}_i$ are not drawn from the prior $f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta})$ conditional on the posterior sample $\boldsymbol{\delta}$, but directly the posterior sample $\boldsymbol{\alpha}_i$ is imputed into the likelihood.

Then custom scalar quantities can be computed, and p-values which compare the fitted posterior-predictive distributions with the actual realizations, as has already been described in section 2.4. If the test statistics are separate for each individual, a histogram of the resulting p-values can be drawn. Good fit is then signalled by a hump-shaped histogram. The p-values could be used for outlier detection, i. e. individuals where some aspects are not fitted well by the model should have striking p-values. If the test statistic summarized all individuals, only one p-value could be reported for the whole model. However, in this chapter we will only use BOT values and score values for outlier detection, because these “test statistics” have a general scope and it is not easy to manufacture summary statistics without knowing the data very well.

### 4.3.3 Special case linear mixed model

In this section the implementation for the normal linear mixed model from section 4.2.3 is detailed.

#### Posterior-predictive samples

The linear mixed model assumes that the observations $y_{i1}, \dots, y_{in_i}$ from one individual $i$ are conditionally independent given the modelled mean vector $\boldsymbol{\mu}_i$ (which is a function of $\boldsymbol{\beta}$ or $\boldsymbol{\xi}$, and $\boldsymbol{\alpha}_i$) and the variance $\sigma^2$:

$$f(\boldsymbol{y}_i \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i) = \prod_{j=1}^{n_i} \mathrm{N}(y_{ij} \,|\, \mu_{ij}, \sigma^2).$$

This is convenient, because `BayesX` can provide us not only posterior samples of $\sigma^2$, but also posterior samples of $\mu_{ij}$.[ii] Therefore the generation of posterior-predictive samples from $f(\boldsymbol{y}_i^* \,|\, \boldsymbol{\xi}, \boldsymbol{\alpha}_i)$ for the goodness-of-fit checks (where $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}_i$ are samples from the full posterior) is reduced to the generation of scalar normal random variates $y_{i1}^*, \dots, y_{in_i}^*$. In particular, we do not need to compute the mean vectors $\boldsymbol{\mu}_i$ resulting from the samples of the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{\alpha}_i$ by ourselves. This can be complicated and error-prone if the fixed effects comprise e. g. basis coefficients of nonlinear spline terms, for which the adequate design matrix would have to be constructed in order to obtain the corresponding contribution to the mean vector.

#### Approximate leave-one-out samples

For the generation of approximate leave-one-out samples, the posterior-predictive sampling approach must be slightly modified. The reason is that the posterior mean samples $\boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i$ produced by `BayesX` were computed with random effects samples $\boldsymbol{\alpha}_i$ from the full conditional distribution $f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta}, \boldsymbol{y}_i)$ in the Gibbs sampler. However, for

[ii] This can be configured with the `regress` method options `predict` and `predictmu`, cf. Belitz, Brezger, Kneib, and Lang (2009b, p. 85).

the approximate leave-one-out cross-validation scheme described on page 100, we need mean samples $\boldsymbol{\mu}_i^*$ deriving from $\boldsymbol{\alpha}_i^* \sim f(\boldsymbol{\alpha}_i \,|\, \boldsymbol{\delta})$. These "prior-predictive" (Marshall and Spiegelhalter 2007, p. 413) random effects samples can be easily produced from normal distributions with variances being the MCMC samples of $\boldsymbol{\delta} = (\delta_1^2, \dots, \delta_q^2)$. Afterwards, we correct the original mean samples $\boldsymbol{\mu}_i$ as follows:

$$\begin{aligned}\boldsymbol{\mu}_i^* &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i^* \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i - \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{Z}_i\boldsymbol{\alpha}_i^* \\ &= \boldsymbol{\mu}_i + \boldsymbol{Z}_i(\boldsymbol{\alpha}_i^* - \boldsymbol{\alpha}_i).\end{aligned}$$

Since `BayesX` allows random intercept and random slope terms, but does not support more complicated random effects terms, the construction of the corresponding design matrices $\boldsymbol{Z}_i$ is straightforward. Finally, the samples $y_{ij}^*$ from the data generating normal distributions with means $\mu_{ij}^*$ and variance $\sigma^2$ are the approximate leave-one-out samples.

### Exact leave-one-out samples

In order to draw exact leave-one-out samples, the `BayesX` sampler must be run $n$ times. If we want to get samples $\boldsymbol{y}_i^*$ for the prediction of the $i$-th individual, given the remaining data $\mathcal{Y}_{\mathcal{N}\setminus\{i\}}$, first we need parameter samples from $f(\boldsymbol{\mu}_i, \sigma^2 \,|\, \mathcal{Y}_{\mathcal{N}\setminus\{i\}})$ – the second step is then again sampling from the normal likelihood. The generation of the parameter samples from the reduced posterior can be achieved by including a weight variable in the data frame, which is 0 if the observation belongs to individual $i$ and 1 else.[iii] Thus, the observations from individual $i$ have no influence on the Bayesian estimation, but the output includes mean samples $\boldsymbol{\mu}_i$ and (of course) regression variance samples $\sigma^2$. Note that here each individual has its own set of $\sigma^2$ samples, while in the approximate sampling scheme all observations share the same variance sample per iteration. This is due to the fact that $\sigma^2$ is part of the non-individual parameter $\boldsymbol{\xi}$, which is only sampled from the full posterior in the approximate sampling scheme.

One potential difficulty is due to the utilized MCMC methods: for valid judgements we need to be sure that the Markov chains have (practically) converged to their stationary distributions, before we use the samples for further computations. Unfortunately, there are no automatic gold-standard checks for MCMC convergence, which could be built into the leave-one-out cross-validation loop. Therefore detailed checks using the original samples from `BayesX` are only straightforward for the posterior-predictive and approximate cross-validation procedures, because they are based on merely a single Markov chain. For the exact cross-validation procedure, the checks can nonetheless be done for the res-

[iii] See Belitz, Brezger, Kneib, and Lang (2009b, p. 63) for the specification of weights in `BayesX`.

ulting parameter samples, which need to be saved in order to later produce the predictive samples.

### Logarithmic scores and BOT values estimation

For the use of the logarithmic score comparing the $i$-th forecaster $F_i$ (having density $f_i$) with the materialized observation $\boldsymbol{y}_i$,

$$LogS(F_i, \boldsymbol{y}_i) = -\log f_i(\boldsymbol{y}_i),$$

the density ordinate $f_i(\boldsymbol{y}_i)$ must be estimated. The general Monte Carlo estimation approach was already described for general models in section 2.3.3. For the model framework described in section 4.2.2, the materialized observation is a time series $\boldsymbol{y}_i$ from an individual $i$. Sampling from the predictive density for this vector proceeds hierarchically, as was already detailed above for the three different sampling schemes: First, parameters $\boldsymbol{\theta}_{i[1]}, \dots, \boldsymbol{\theta}_{i[m]}$ are drawn from the full or reduced posterior, where $\boldsymbol{\theta}_i$ comprises $(\boldsymbol{\mu}_i, \sigma^2)$. Second, the samples $\boldsymbol{y}_{i[1]}, \dots, \boldsymbol{y}_{i[m]}$ will be drawn from the $m$ resulting conditional densities $f_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{i[1]}), \dots, f_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{i[m]})$.

While the marginal density ordinate $f_i(\boldsymbol{y}_i)$ is unknown, the conditional density ordinate $f_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_i)$ is known for all $\boldsymbol{\theta}_i$:

$$f_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \mathrm{N}(y_{ij} \,|\, \mu_{ij}, \sigma^2).$$

Thus, given the model parameter samples $\boldsymbol{\theta}_{i[k]}$, we can again use the Monte Carlo estimate (2.3.4) which is $\hat{f}_i(\boldsymbol{y}_i) = \frac{1}{m}\sum_{k=1}^{m} f(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{i[k]})$, and impute it into the logarithmic score formula. So the full estimate for the logarithmic score comparing the $i$-th forecaster $F_i$ with the materialized observation $\boldsymbol{y}_i$ is

$$\widehat{LogS}(F_i, \boldsymbol{y}_i) = -\log \hat{f}_i(\boldsymbol{y}_i) = \log(m) - \log \sum_{k=1}^{m} f_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{i[k]}).$$

The estimates $-\log \hat{f}_1(\boldsymbol{y}_1), \dots, -\log \hat{f}_n(\boldsymbol{y}_n)$ can then be averaged to obtain the mean log score of the model. Note that this mean log score is not identical to the average of the log scores of the individual scalar observations $y_{ij}$, since

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_i(\boldsymbol{y}_i) &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{m} \sum_{k=1}^{m} \prod_{j=1}^{n_i} f(y_{ij} \,|\, \mu_{ij[k]}, \sigma^2_{i[k]}) \\ &\neq \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \log \prod_{j=1}^{n_i} \frac{1}{m} \sum_{k=1}^{m} f(y_{ij} \,|\, \mu_{ij[k]}, \sigma^2_{i[k]}) = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log \hat{f}_{ij}(y_{ij}).\end{aligned}$$

In order to compute the BOT estimates $\widehat{BOT}(f_i, \boldsymbol{y}_i)$ $(i = 1, \dots, n)$ from (2.2.3), we do not only need the density ordinates $z_{\boldsymbol{y}_i} = \hat{f}_i(\boldsymbol{y}_i)$ of the materialized observations $\boldsymbol{y}_i$ under the respective forecast densities $f_i$, but we also need the density ordinates $z_{i[l]} = \hat{f}_i(\boldsymbol{y}_{i[l]})$ of the forecast samples $\boldsymbol{y}_{i[l]}$, $l = 1, \dots, m$. These forecast samples are available anyway as they are necessary for the estimation of the energy score. We again use the straightforward Monte Carlo estimates

$$\hat{f}_i(\boldsymbol{y}_{i[l]}) = \frac{1}{m} \sum_{k=1}^{m} f_i(\boldsymbol{y}_{i[l]} \,|\, \boldsymbol{\theta}_{i[k]}), \tag{4.3.2}$$

which are computed for all forecast samples. The BOT estimate is then the fraction of estimated forecast ordinates being smaller than the estimated observation ordinate:

$$\widehat{BOT}(f_i, \boldsymbol{y}_i) = \frac{1}{m} \sum_{l=1}^{m} \mathbb{I}_{[z_{i[l]}, +\infty)}(z_{\boldsymbol{y}_i}).$$

## 4.4 Simulation study

In order to test the proposed predictive sampling schemes in a situation where the true linear mixed model underlying the data is known, we do a small simulation study in this section. The model which generates the test data set is described in the following.

The data set comprises $n = 40$ individuals with one binary covariate (e. g. sex), which is $z_i = 0$ (male) for the first half and $z_i = 1$ (female) for the second half. The number of observations, $n_i$, is drawn from the discrete uniform distribution $\mathrm{U}\{3, \dots, 10\}$, iid for all individuals $i = 1, \dots, n$. The observation times $t_{ij}$ are then generated iid from the continuous uniform distribution $\mathrm{U}(0, 10)$, $j = 1, \dots, n_i$. The response values are generated from normal distributions with mean

$$\mu_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 z_i + \alpha_{i1} + \alpha_{i2} t_{ij} \tag{4.4.1}$$

for time $j$ of individual $i$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ collects the fixed population effects (intercept, slope, baseline difference for females) and $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2})'$ collects the random effects for individual $i$ (baseline and slope differences). So we have $p = 3$, $q = 2$,

$$\boldsymbol{Z}_i = (\boldsymbol{1}_{n_i} \,|\, \boldsymbol{t}_i) \quad \text{and} \quad \boldsymbol{X}_i = (\boldsymbol{Z}_i \,|\, z_i \boldsymbol{1}_{n_i}),$$

where $\boldsymbol{t}_i = (t_{i1}, \dots, t_{in_i})'$.

We set $\boldsymbol{\beta} = (2, 7, 10)'$ as the fixed effects, and $\sigma^2 = 4$ as the noise variance. The random effects are independent Gaussian draws with variance $\delta_1^2 = \delta_2^2 = 9$. The resulting data set is graphed in Figure 4.3.

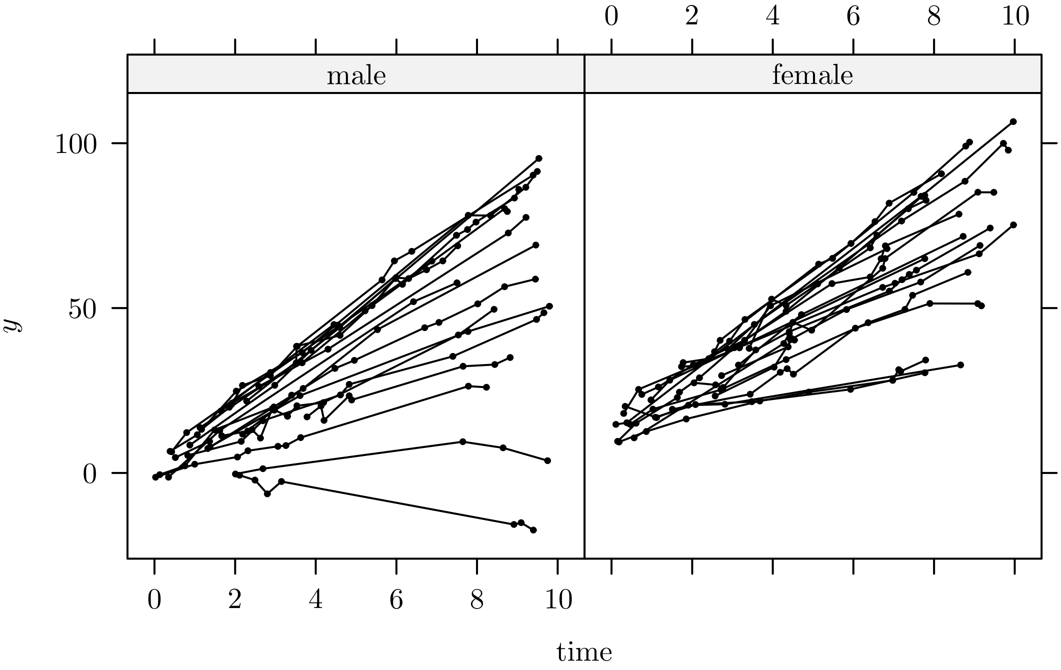


***Figure 4.3*** – *Simulated longitudinal data set.*

### 4.4.1 Performance of the correct model

First we want to check how the correct model performs. We choose the (default) hyperparameters $a = b = c = d = 0.001$ for the prior of the variance parameters $\sigma^2$ and $\boldsymbol{\delta}$.

For generating 200 000 iterations from which every 20-th sample was saved, `BayesX` needed only 64 seconds. From the traceplot of the saturated deviance (cf. appendix A.2) in Figure 4.4 we can assume that after 40 000 iterations the Markov chain has practically converged to its stationary posterior distribution. We will thus discard the previous saved samples as the burn-in phase, and work with the resulting 8000 thinned-out samples.

While the generation of the approximate cross-validation model parameter samples is done in 2 seconds, the generation of the exact equivalents takes 54 minutes – so the approximate approach is 1791 times faster! This is because we have to run `BayesX` 40 times again (with the same MCMC parameters as for the full data run) to get the exact results, but can use the already existing samples obtained from the full data to get the approximate results. We checked the convergence of the reduced data Markov chains in the exact cross-validation procedure by looking at traceplots of some means and variance samples. A burn-in of 2000 for the thinned-out samples appeared adequate here too. The production of the

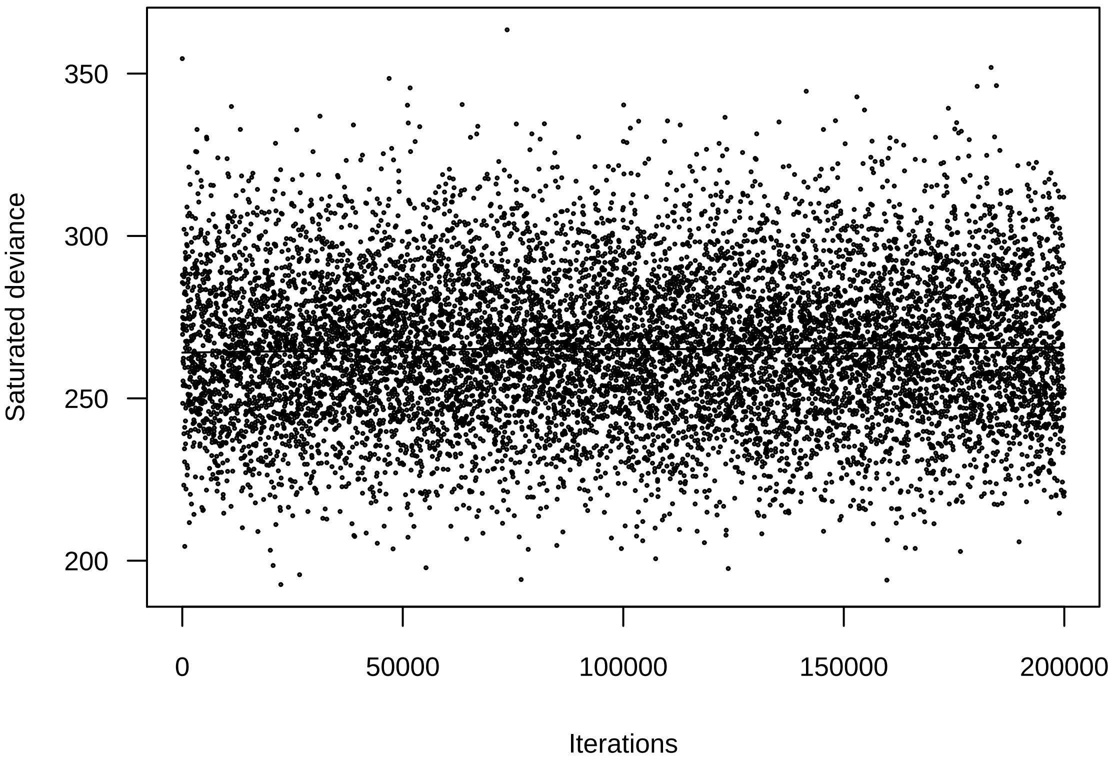
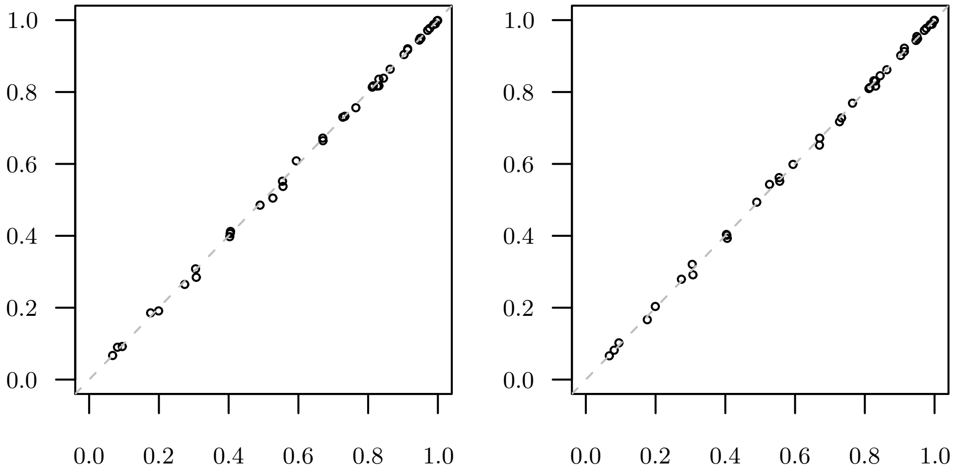


***Figure 4.4*** – *Traceplot of the saturated deviance.*

resulting forecast samples worked then the same way for all three procedures (posterior-predictive, approximate and exact cross-validation), conditional on the respective model parameter samples (cf. section 4.3.3).

The estimation of the BOT values raised computational questions, which already emerged with the estimation of the posterior-predictive BOT values: As described on page 105, we first used all $m = 8000$ model parameter samples. This led to a required computing time of 2430 seconds or 40 minutes, even after a 5-fold acceleration of the slower original R-code using highly optimized `C++`-code. The problem is that the complexity of the algorithm is $O(r \cdot m \cdot \sum_{i=1}^{n} n_i)$, if $r$ is the number of used model parameter samples $\{\boldsymbol{\theta}_{i[k]}\}_{k=1}^{r}$ and $m$ is the number of used predictive samples $\{\boldsymbol{y}_{i[l]}\}_{l=1}^{m}$. Since the effort was inconvenient for regular practical use, we tried to use $r \ll m$ parameter samples. Fortunately we found that a subset of $r = 200$ randomly chosen samples yields very good approximations to the BOT values obtained from all $m = 8000$ samples, while requiring proportionally less computing time – only 56 seconds in our example. The results are also stable with regard to the choice of subset. The full sample BOT values are compared with two approximations in Figure 4.5. Given these promising results, we will always use $r = 200$ randomly chosen model parameter samples for the BOT values estimation from now on.

***Figure 4.5*** – *Comparison of full samples BOT estimates (8000 samples) with two approximations, resulting from different subsets of size 200. Obviously the differences between full and approximate values are negligible, with the maximum deviances being* 0.023 *and* 0.018*, and the mean deviances being* 0.006 *and* 0.005 *for the two subsets, respectively.*



**(a)** First subset

**(b)** Second subset

The resulting BOT and scalar-PIT histograms of the posterior-predictive, exact and approximate cross-validation predictions are shown in Figure 4.6. The clearest picture is given by the histograms for the posterior-predictive forecasts. The PIT histogram in panel (d) has the typical hump-shaped form, meaning that the posterior-predictive forecasts are overdispersed compared to the original data. The corresponding BOT histogram in panel (a) is heavily left-skewed, which fits the picture of overdispersed predictive samples. Of course, this overdispersion is in fact desirable here, because it means that the posterior-predictive distributions are centered around the original observations. The BOT histogram for the exact leave-one-out forecasts in panel (c) fulfills the expectations quite well – no large deviation from uniformity is visible. The corresponding PIT histogram in panel (f), obtained from the $n$-fold cross-validation of the individual scalar observations, is more difficult to interpret. At least no clear over- or underdispersion can be diagnosed. The approximate counterpart PIT histogram in panel (e) is very similar. The BOT histogram in panel (b) shows only slightly worse calibration. Overall, the approximate histograms are good surrogates for the exact histograms.

The approximation of the exact cross-validation logarithmic scores, energy scores and BOT values with either the posterior-predictive or the proposed approximate sampling scheme is visualized in Figure 4.7. Both the exact logarithmic scores in panel (d) and the energy scores in panel (e) are very well approximated by the proposed sampling scheme. This is in contrast to the posterior-predictive scores, which are systematically much lower than the exact scores. While at least a linear correlation between the exact leave-one-out and posterior-predictive logarithmic scores is seen in panel (a), the posterior-predictive energy scores are heavily shrunk towards 0 in panel (b). The approximation of individual BOT values with the proposed fast sampling scheme seems to be more difficult. In panel (f) much larger differences than e. g. in Figure 4.5 are reported. This explains that the resulting BOT histograms in Figure 4.6 are noticeably different. Panel (c) shows that the posterior-predictive BOT values are almost always larger than the exact BOT values in this example.

In Table 4.1 the mean scores are presented. A small amount of conservativeness of the approximate mean scores can be seen, as they are lower than the respective exact mean scores. However, the differences are rather small and the order of magnitude is preserved. By contrast, the posterior-predictive mean scores are much lower than the exact mean scores, with the posterior-predictive energy score being almost an order of magnitude below the exact energy score.

***Table 4.1*** – *Mean energy and logarithmic scores under the exact and approximate leave-one-out and posterior-predictive sampling schemes.*

| Scoring rule | exact | approximate | posterior-predictive |
|---|---|---|---|
| ES | 23.96 | 23.27 | 3.03 |
| log-score | 17.81 | 17.65 | 13.73 |

### 4.4.2 Comparison with other models

Now we want to see how sensitive the model assessment is to the omission of important features of the true model. We consider three (partially) wrong models.

For the first model, we omit the binary covariate (the term $\beta_3 z_i$ in formula (4.4.1)), while for the second model we omit the random slope ($\alpha_{i2} t_{ij}$) in the specification of the linear predictor. Both the covariate and the random slope are omitted for model 3. So model 1 misses a fixed effect, model 2 misses a random effect and model 3 misses both a fixed and a random effect of the true model. This time we discard the burn-in of 40 000 iterations directly in `BayesX` and keep the other MCMC parameters from the correct model sampling in section 4.4.1. The required computing takes 73, 58 and 56 seconds for the three models, respectively.

In order to compare the goodness-of-fit of all four models (the correct model plus the three wrong models), we look at the mean deviance and posterior-predictive energy and log-scores in Table 4.2. It is instructive that the wrong model 1 has the best bit of all models, with respect to all three fit criteria, although the correct model is only slightly worse. This is due to the random intercept term in model 1, which absorbs the difference between male and female baseline levels. If we had omitted a time-varying covariate, this would not have worked, but the binary covariate was time-constant in the example. Model 2 has clearly a worse fit to the data, with the deviance difference being much smaller than the scores differences. Model 3 is slightly better than model 2, but the gap between the two models is not large, for the same reason that the correct model and model 1 are very close. For both model pairs, the model without the fixed binary covariate effect has a better fit.

***Table 4.2*** – *Posterior-predictive mean energy and logarithmic scores as well as the posterior expected saturated deviances of the correct and the three wrong models.*

| Fit criterion | Correct model | Model 1 | Model 2 | Model 3 |
|---|---:|---:|---:|---:|
| ES | 3.03 | 2.98 | 13.49 | 13.48 |
| log-score | 13.73 | 13.69 | 23.67 | 23.64 |
| Deviance | 265.88 | 265.69 | 266.02 | 265.98 |

The posterior-predictive BOT and PIT histograms are shown in Figure 4.8. Here the goodness-of-fit ranking is more difficult. The model 1 BOT histogram in panel (a) is very similar to the correct model's histogram in panel (a) on page 113. If we compare the model 2 and model 3 BOT histograms in panels (b) and (c), we rather come to the conclusion that they are fitting the data better than the correct model, because they show a larger frequency of high BOT values than the correct model's BOT histogram. The model 1 PIT histogram in panel (d) is similar and slightly more heavily hump-shaped than the original model's PIT histogram in panel (d) on page 113. Ordering the goodness-of-fit of model 2 and model 3 and the fit of the correct model using the PIT histograms in panels panels (e) and (f) is very difficult.

For the three wrong models, we also produced cross-validation parameter samples using the exact leave-one-out sampling scheme, which required 2676, 2929 and 2532 seconds. Again the approximate sampling was much faster with only 1 additional second being required for each of the three models.

The exact cross-validation BOT and PIT histograms are shown in Figure 4.9. The model 1 BOT histogram in panel (a) is almost as close to a uniform histogram as the correct model's counterpart in panel (c) on page 113. The difference of the two other models is obvious in their cross-validation BOT histograms in panels (b) and (c): they are much more distinct from uniform histograms. The scalar PIT histograms do not give a comparably clear picture, with all histograms differing from each other and from the correct model's PIT histogram in panel (f) on page 113. So the BOT histograms seem to be more useful here.

The mean cross-validation scores and DIC values for all four models are presented in Table 4.3 (see appendix A.2 for the DIC definition). The exact energy and logarithmic scores agree that the correct model has the best predictive performance, followed by models 1, 2 and 3. This ranking is perfectly preserved by the approximate logarithmic scores. The approximate energy scores rank model 2 better than model 1. By contrast, the DIC ranks model 2 and model 3 best, and sends the correct model down to the third place with a large difference of the DIC value.

***Table 4.3*** – *Mean energy and logarithmic scores for the cross-validated prediction of the simulated data for the correct model and the three wrong models, under the exact and approximate sampling schemes. The DIC values based on the saturated deviance samples reported by `BayesX` are also shown.*

| Model criterion | Scheme | Correct model | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|---|
| ES | exact | 23.96 | 25.95 | 26.29 | 27.45 |
| | approximate | 23.27 | 25.20 | 25.05 | 26.61 |
| log-score | exact | 17.81 | 18.60 | 25.83 | 25.96 |
| | approximate | 17.65 | 18.29 | 25.18 | 25.34 |
| DIC | | 337.10 | 344.06 | 306.61 | 306.78 |

### 4.4.3 Results

The experiments with the simulated data set have emphasized that the posterior-predictive results are only useful for a goodness-of-fit assessment of the models in question. In doing so, the posterior-predictive scores should be preferred over PIT and BOT histograms, due to their easier interpretability. The distinction of goodness-of-fit assessment on the one hand and predictive assessment on the other hand is a very important point because an equally good fit does not imply equally good prediction of new data. For example, the wrong model 1 fitted the data equally well as the correct model, but was of course outperformed by the correct model in the prediction of the left-out data.

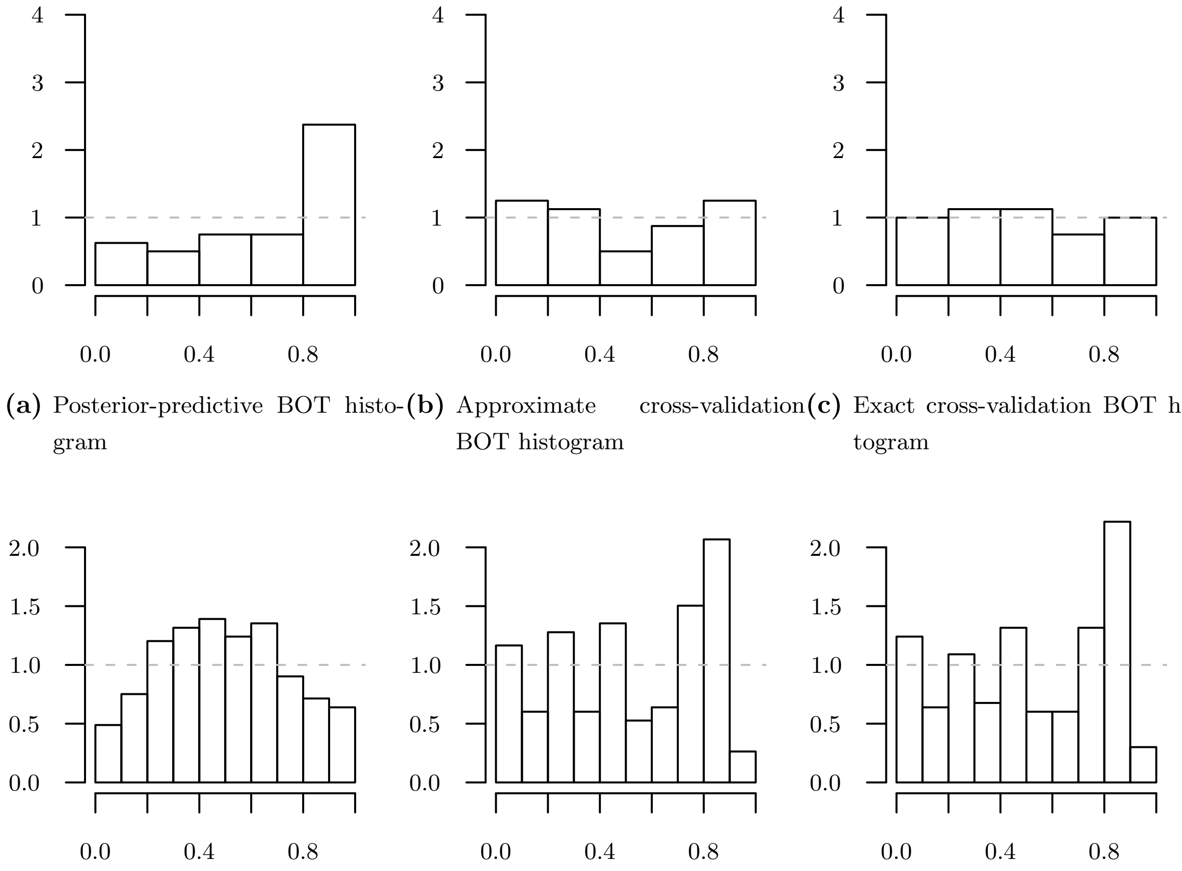
The deviance and DIC measures appear to be less useful than proper scoring rules: The deviance showed only small differences in the goodness-of-fit assessment, and the DIC yielded a wrong model ranking (the correct model was not the best model) in the predictive performance assessment.

The approximate sampling scheme worked very well for the log-scores in this example.

The approximations worked also well for energy scores. Only the approximation of the BOT values and corresponding histograms seems to be more difficult. Overall, the trade-off between computational efficiency and good approximation of the leave-one-out results seems to be fine.

However, these results cannot be generalized to larger applications, because we have only conducted a very small simulation study here, both with respect to the number of individuals/observations and the use of only one simulated data set. A more serious simulation study would need to be done in large scale with replications of data sets.

***Figure 4.6*** – *Multivariate BOT (upper row) and scalar PIT (lower row) histograms for calibration assessment of the leave-one-out prediction in the correct random effects model. The predictive distributions were estimated with the posterior-predictive (left column), approximate (middle column) and exact (right column) cross-validation sampling schemes. Only 5 bins were used for the BOT histograms because of the small sample size of $n = 40$. On the other hand, 10 bins were used for the PIT histograms, where the sample sizes are larger ($\sum_{i=1}^{n} n_i = 266$).*



**(a)** Posterior-predictive BOT histogram

**(b)** Approximate cross-validation BOT histogram

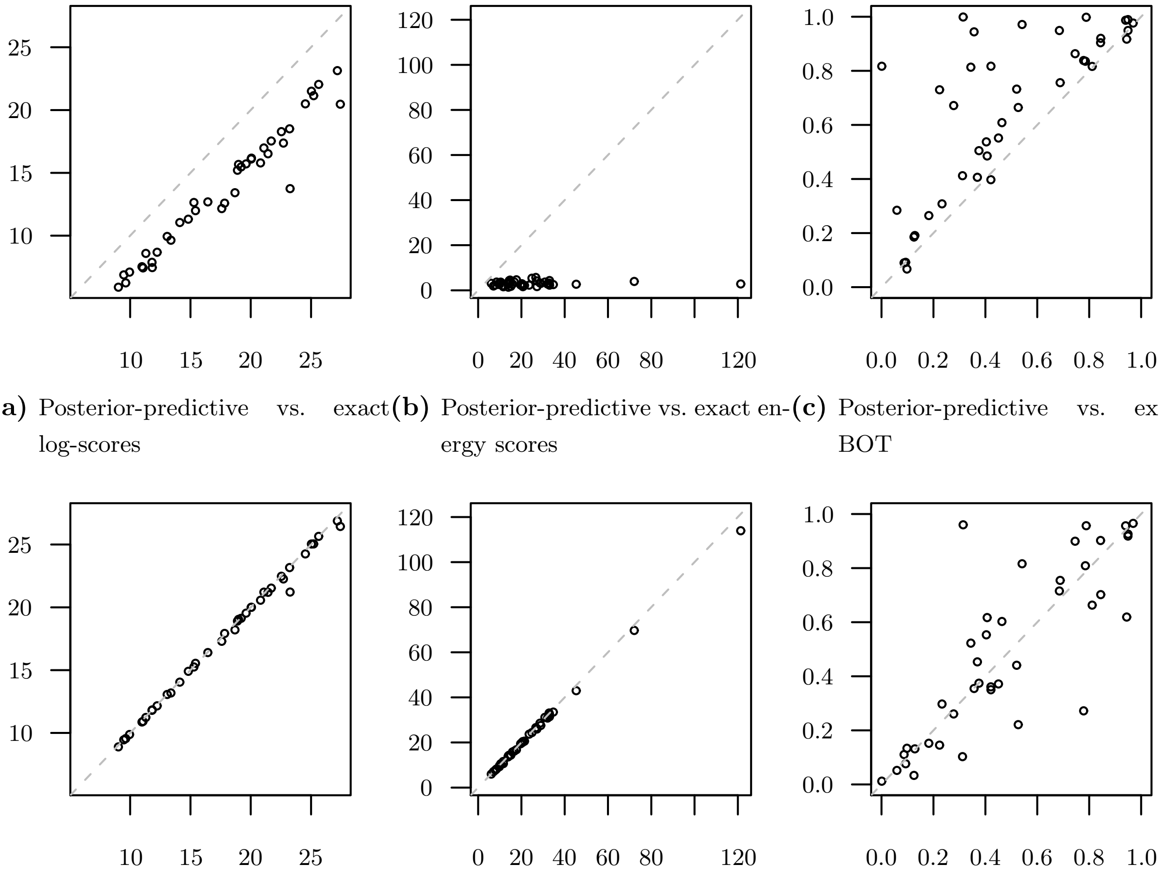
**(c)** Exact cross-validation BOT histogram

**(d)** Posterior-predictive PIT histogram

**(e)** Approximate cross-validation PIT histogram

**(f)** Exact cross-validation PIT histogram

***Figure 4.7*** – *Comparison of the approximation of the exact cross-validation logarithmic scores (left column), energy scores (middle column) and BOT values (right column) with either the posterior-predictive (upper row) or the proposed approximate sampling scheme (lower row). The exact values are the x-axis coordinates, while the approximate values are the y-axis coordinates.*



**(a)** Posterior-predictive vs. exact log-scores

**(b)** Posterior-predictive vs. exact energy scores

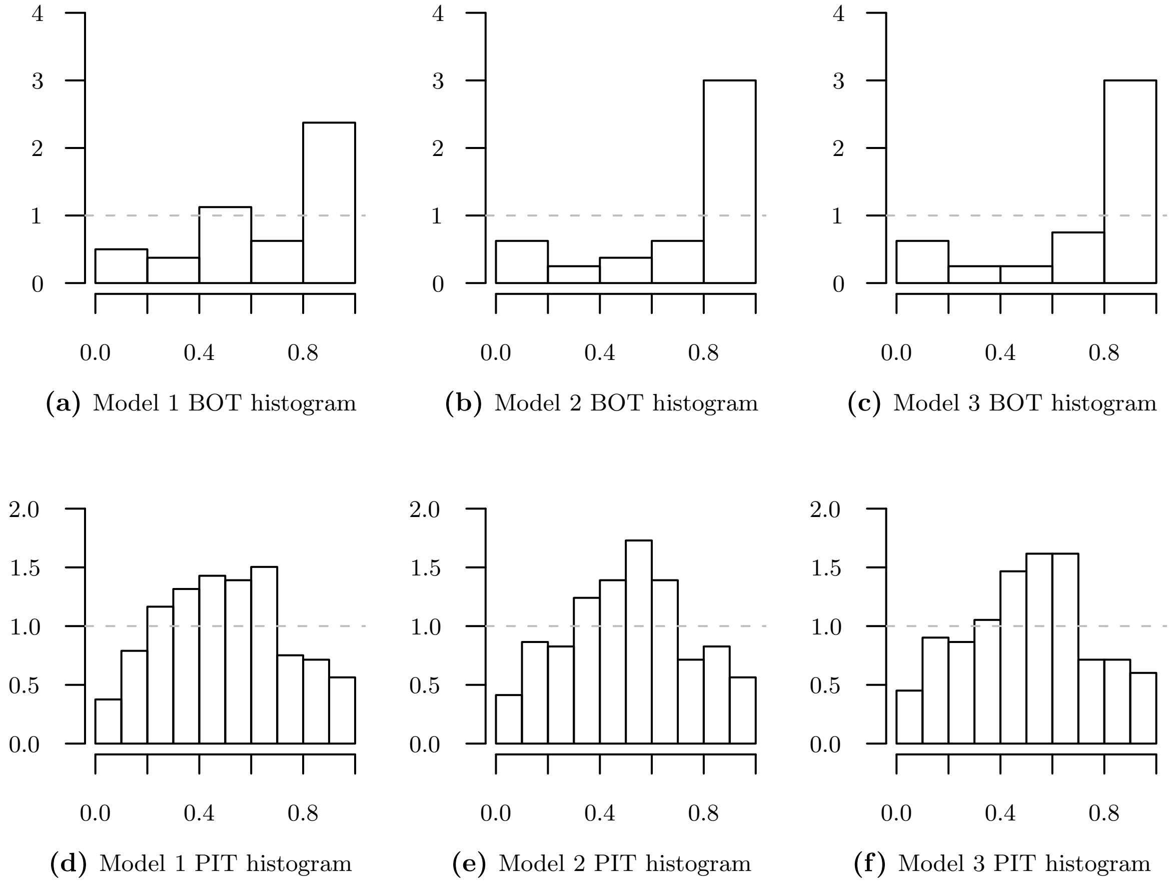
**(c)** Posterior-predictive vs. exact BOT

**(d)** Approximate vs. exact log-scores

**(e)** Approximate vs. exact energy scores

**(f)** Approximate vs. exact BOT

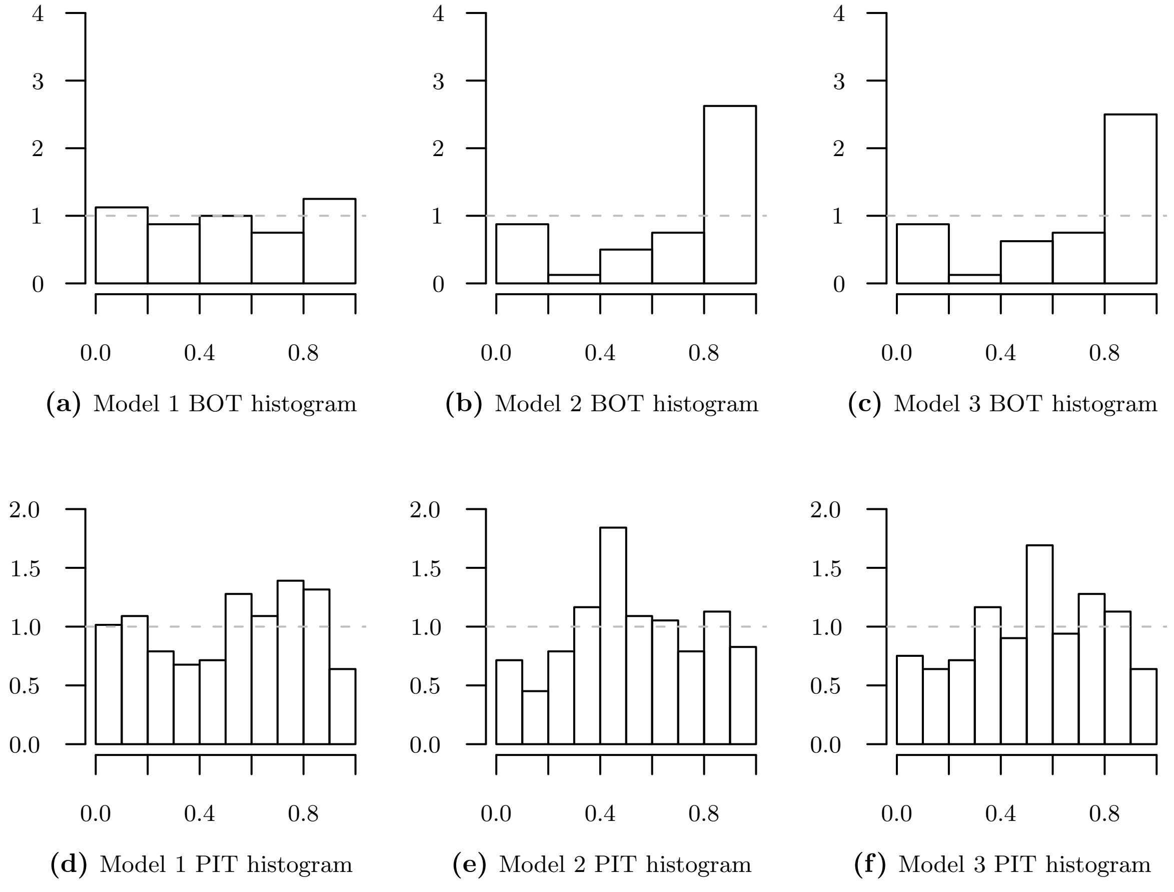
***Figure 4.8*** – *Posterior-predictive multivariate BOT (upper row) and scalar PIT (lower row) histograms for goodness-of-fit assessment of the three wrong random effects models (columns).*



**(a)** Model 1 BOT histogram **(b)** Model 2 BOT histogram **(c)** Model 3 BOT histogram

**(d)** Model 1 PIT histogram **(e)** Model 2 PIT histogram **(f)** Model 3 PIT histogram

***Figure 4.9*** – *Exact cross-validation BOT (upper row) and PIT (lower row) histograms for leave-one-out calibration assessment of the three wrong random effects models (columns).*



**(a)** Model 1 BOT histogram **(b)** Model 2 BOT histogram **(c)** Model 3 BOT histogram

**(d)** Model 1 PIT histogram **(e)** Model 2 PIT histogram **(f)** Model 3 PIT histogram

## 4.5 CD4 data

The CD4 data set has been compiled by Andreas Bayerstadler from the Multicenter AIDS cohort study (MACS) Public Data Set Release P17, which can be ordered via the Internet[iv]. The MACS study is a long-term prospective cohort study of homosexual men recruited at four study centers in the US, which started in April 1984. The cut-off date for the used release P17 is 1st October 2004. Biannually, the participants were tested for human immunodeficiency virus (HIV) positivity, to estimate the date of HIV seroconversion. Moreover, detailed questionnaires, physical examinations and other laboratory tests were carried out. A detailed description of the MACS study can be found elsewhere (Kaslow et al. 1987).

Our data set comprises $n = 574$ patients who appeared for the biannual interviews between $\min n_i = 1$ and $\max n_i = 41$ times, which leads to a total of $\sum n_i = 10\,606$ individual observations. The (quasi-)continuous response variable is the number of the T helper cells expressing the surface protein CD4 in a fixed blood volume. These special white blood cell are therefore called CD4 positive cells, or Leu-3 cells, which is why the variable is named `LEU3N`. See Janeway et al. (1988) for an early review. Low CD4 lymphocyte counts are associated with increased risk of progression to AIDS in HIV infected persons (Lee et al. 1991). Therefore we are interested in modelling the individual CD4 counts trajectories, conditional on the covariates listed in Table 4.4.

Typical trajectories are graphed for a random subset of 18 patients in Figure 4.10. We see that while many seroconverters suffer from a decline in the number of CD4 cells, there is even a patient (ID 5829) with steadily increasing CD4 counts *after* his HIV infection. This might also be due to the availability of more and more anti-HIV active ingredients in the mid-1990s. Patient ID 9963 supports this hypothesis, with a surge in CD4 counts after calendar time 14, which corresponds to the year 1998, when already 13 active ingredients were internationally approved.[v] Large inter-patient variability is observed, both regarding the absolute level of the trajectories and the shape of the time series. It is clear that the covariates from Table 4.4 will not be able to explain most of this variability, but that there are contributing unobserved covariates. We will thus use random effect models to adjust for these influences.

In section 4.5.1 we will first do a complete case analysis of the data. This allows an exact leave-one-out cross-validation assessment of three different models. The results are compared to those from the proposed approximate cross-validation scheme. Section 4.5.2 fits six different models to the whole CD4 data set, with the form of the time effect being

[iv] MACS Web Site: `http://www.statepi.jhsph.edu`

[v] See e. g. `http://www.vfa.de/de/forschung/txt/aids-medikamente-klassen.html` for an overview of active ingredients classes.

| Covariate | Description |
|---|---|
| `DATE` | Estimated date of seroconversion (in years after 1984-01-01) |
| `TIME` | Time of the visit in years after estimated seroconversion |
| `PSSCO` | Psychological score (quasi-continuous) |
| `PACKS` | Number of cigarette packs smoked per day (none, up to half, one, two or more packs per day) |
| `SMOKE` | Number of cigarette packs smoked per day when smoked most (none, up to half, one, two or more packs per day) |
| `NSEX` | Number of sexual partners in the last six months (none, one or more) |
| `DRUGS` | Injection of recreational drugs (binary) |

***Table 4.4*** – *Description of CD4 data set covariates. The sum of `DATE` and `TIME` recovers the calendar time of the visit (in years after 1984-01-01). The psychological score is an average of 20 individual answers coding the frequency of rare (1), some (2), occasional (3), or frequent (4) negative feelings (e. g. how often one felt lonely during the last half year).*

varied. Finally, in section 4.5.3 we include the significant covariates from section 4.5.1 into the best model from the approximate cross-validation in section 4.5.2.

### 4.5.1 Complete case analysis

First we want to include all covariates in the model selection. We therefore discard all observations where any covariate value is missing, and obtain a complete data set with a total of 1040 data points from 111 individuals ($1 \leq n_i \leq 31$). The smaller dimensions will give us the possibility to compare exact and approximate cross-validation results, which would not be possible with the original $n = 574$.

#### Model fitting

The first model includes the three categorical covariates `NSEX`, `DRUGS`, `PACKS` (using appropriate binary dummy variables) and the continuous score `PSSCO`. In addition, a “hockey-stick” assumption is made for the effect of time since seroconversion, where the change of slope can appear at the seroconversion (the origin of the variable `TIME`). In order to include our hypothesis of better medical treatment in the mid-1990s into the analysis, we also allow a change at the beginning of the year 1995 or later if the seroconversion had not taken place yet. These time effects and the intercept are specified as random effects, to adjust for “random” heterogeneity between the individuals. The `BayesX` model formula

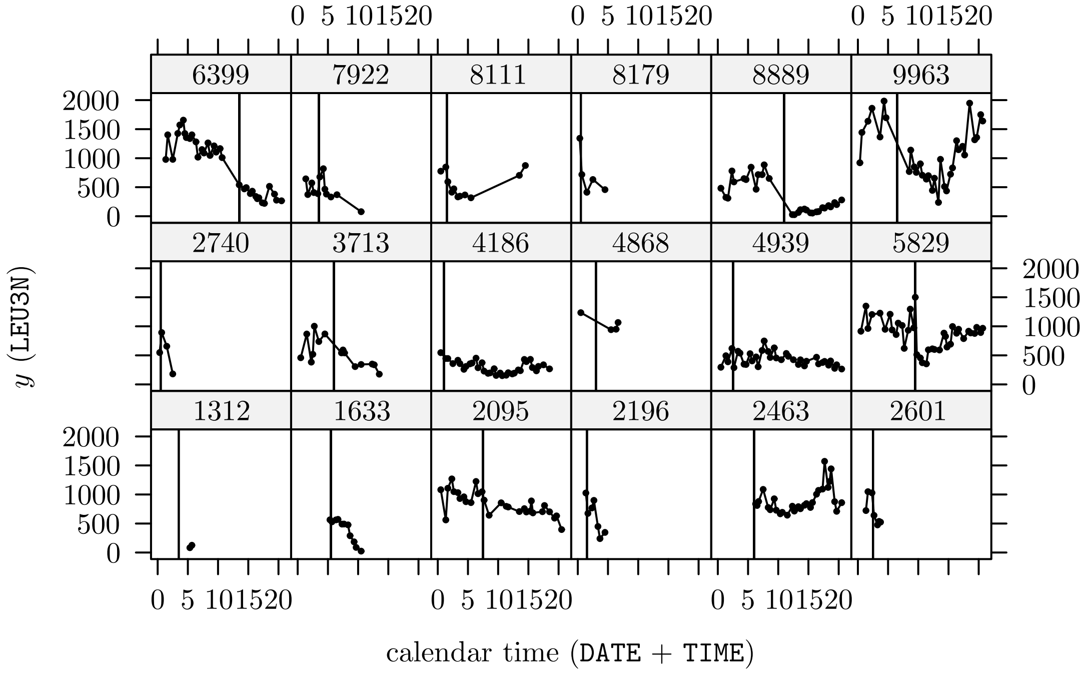


***Figure 4.10*** – *CD4 cell counts trajectories for a random subset of 18 participants, whose IDs are in the headings of the panels. The estimated dates of seroconversion (variable* `DATE`*) are marked by vertical lines.*

has then the following form:

$$\begin{aligned}\texttt{LEU3N} = {} & \texttt{PSSCO} + \texttt{NSEXone} + \texttt{NSEXmore} + \texttt{DRUGSyes} + \\ & \texttt{PACKShalf} + \texttt{PACKSone} + \texttt{PACKStwo} + \texttt{PACKSmore} + \\ & \texttt{CASEID}(\textit{random}) + \texttt{TIME} * \texttt{CASEID}(\textit{random}) + \\ & \texttt{TIMEpos} * \texttt{CASEID}(\textit{random}) + \texttt{TIMEposLate} * \texttt{CASEID}(\textit{random}),\end{aligned}$$

where we have defined the covariates `TIMEpos` as the positive part of `TIME`. The design variable encoding the possible second change is

$$\texttt{TIMEposLate} := (\texttt{TIME} - \max\{0, 11 - \texttt{DATE}\})_+,$$

because 1995 is 11 years after 1984 which is the origin of the variable `DATE`.

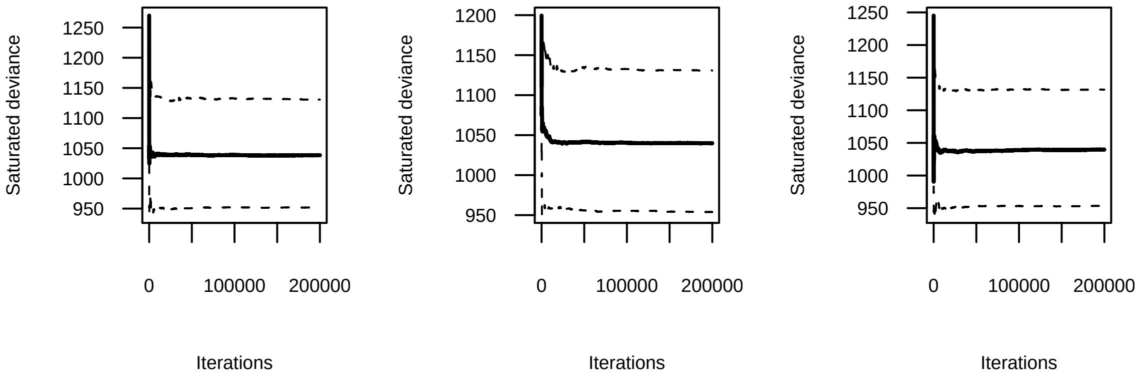
The second model only includes a random intercept to adjust for different baseline CD4 levels, and uses fixed time effects as in the first model. The other covariates are inherited:

$$\begin{aligned}\texttt{LEU3N} = \texttt{PSSCO} + \texttt{NSEXone} + \texttt{NSEXmore} + \texttt{DRUGSyes} +&\\ \texttt{PACKShalf} + \texttt{PACKSone} + \texttt{PACKStwo} + \texttt{PACKSmore} +&\\ \texttt{TIME} + \texttt{TIMEpos} + \texttt{TIMEposLate} + \texttt{CASEID}(random).&\end{aligned}$$

The third model is more parsimonious with restriction to the time-constant covariates SMOKE and DRUGS, a random intercept and a random slope for the time since seroconversion:

$$\begin{aligned}\texttt{LEU3N} = \texttt{DRUGSyes} + \texttt{SMOKEhalf} + \texttt{SMOKEone} + \texttt{SMOKEtwo} + \texttt{SMOKEmore}+&\\ \texttt{CASEID}(random) + \texttt{TIME} * \texttt{CASEID}(random).&\end{aligned}$$

***Figure 4.11*** – *Cumulative saturated deviance quantile plots (median, lower and upper 2.5% quantiles) for the three models.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

We produced Markov chains of length 200 000 and saved every 20-th iteration for the three models, which required 268, 229 and 201 seconds, respectively. The cumulative quantile plots for the saturated deviance in Figure 4.11 suggest that a burn-in of 20 000 for the raw samples, or 1000 for the saved samples, is sufficient. The DIC values (which are estimated from the saturated deviance samples) are 1209.04, 1138.57 and 1179.04, respectively. So model 2 would be preferred by DIC, followed by model 3 and model 1. Note that the DIC values were computed from the whole saved samples chain, including the burn-in which we have discarded later.

In Figure 4.12 the estimated posterior means from the three models are plotted for the patients from Figure 4.10 which are still present in the data set with complete observations. Note that e. g. for patient ID 4186 there is only one observation left from the original 34. Also for ID 8889, the last observations from the original data set are missing, and perhaps this leads to model 1 and model 2 fits without the suspected late upward trend. For

ID 9963, we see a late upward trend in the model 1 fit. Nevertheless, the posterior mean fixed effect part of change of slope after 1995 (`TIMEposLate`) is positive for both models (69.06 and 57.27).

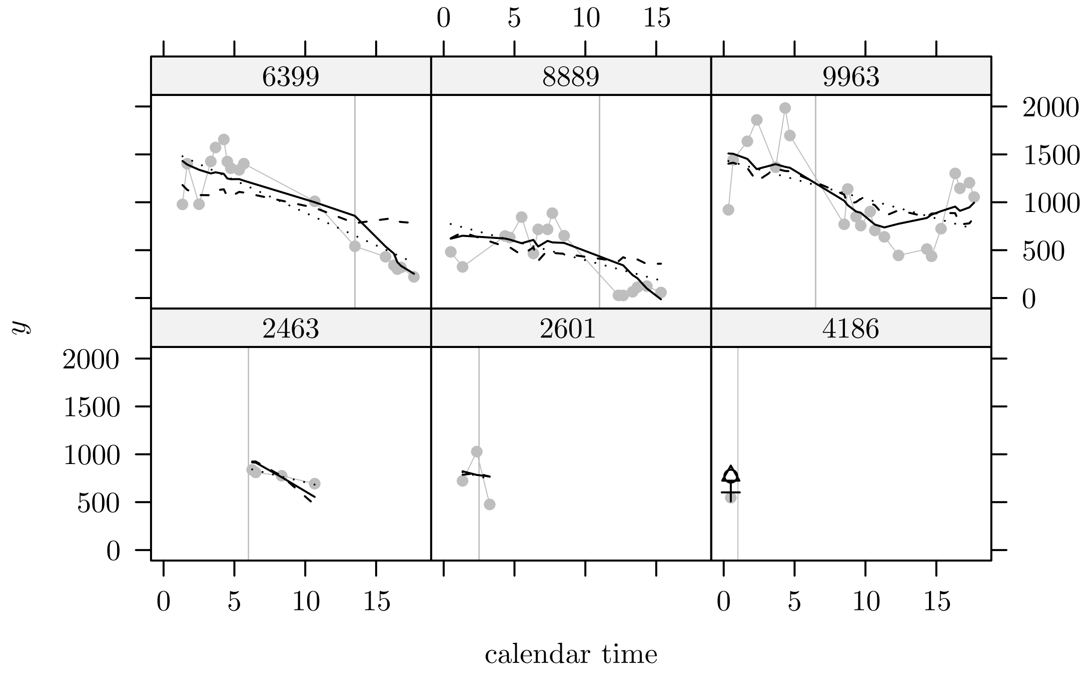


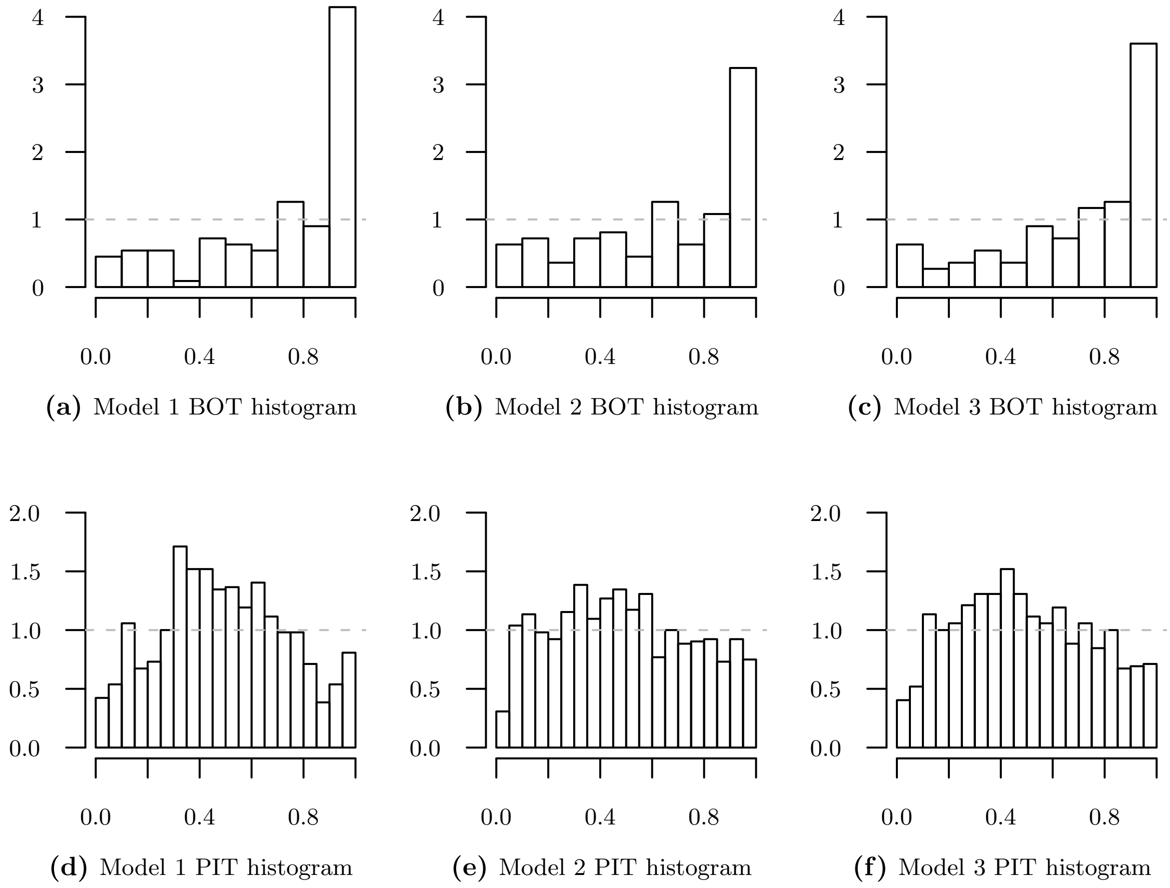
*__Figure 4.12__ – Model fits (estimated posterior means) for the patients from Figure 4.3 which are still present in the data set with complete observations. The three models are discerned by line (or point) type: model 1 (——, ○), model 2 (- - -, △) and model 3 (........, +). The original data is plotted in gray.*

### Goodness-of-fit assessment

Overall the three model fits do not differ much. In order to assess the goodness-of-fits, we look at posterior-predictive BOT and PIT histograms in Figure 4.13. Model 1 and model 3 seem to have a better fit to the given data than model 2: Their BOT histograms in panels (a) and (c) are more left-skewed than the Model 2 BOT histogram in panel (b). Also the PIT histogram for the individual scalar observations in panel (e) shows that model 2 generates more PIT values above 0.8 than both other models. Model 1 looks best here.

Outlying individuals can be characterized by a small BOT value, because that means

***Figure 4.13*** – *Multivariate BOT (upper row) and scalar PIT (lower row) histograms for goodness-of-fit assessment of the three random effects models (columns).*



**(a)** Model 1 BOT histogram

**(b)** Model 2 BOT histogram

**(c)** Model 3 BOT histogram

**(d)** Model 1 PIT histogram

**(e)** Model 2 PIT histogram

**(f)** Model 3 PIT histogram

that the probability of observing a smaller posterior-predictive density ordinate is small. Thus, the materialized (multivariate) observation is in a low-density region of the posterior-predictive distribution. As we expect the posterior-predictive distributions to center around the known observations and to assign high density to their neighborhood, that is indeed an argument for an outlying observation. We show the 7 observations having a BOT value smaller than 0.05 in at least one of the three models in Figure 4.14.

For ID 5587, the fit from model 1 is obviously better than from both other models. This is nicely reflected by the large BOT value (0.409) while the both other models have BOT values below 0.1. IDs 3635, 4437, 5419 are not fitted well by all models: For the latter two IDs, an individual outlying first observation at the beginning of the time series (with corresponding scalar PIT value 1 for all models) is probably the reason for the relatively small posterior-predictive density ordinates. For ID 3635, some points are distant from the mean fits, but the cause for the low BOT values is not so obvious from the plot. Note however that we have only visualized the means of the posterior-predictive distributions for the time points, but not quantiles or the full forms of the respective densities. The BOT assessment takes account of the full distribution, and not only the mean. So a small BOT value could also be due to an underdispersed predictive distribution, although the recorded observation lies close to the predictive mean.

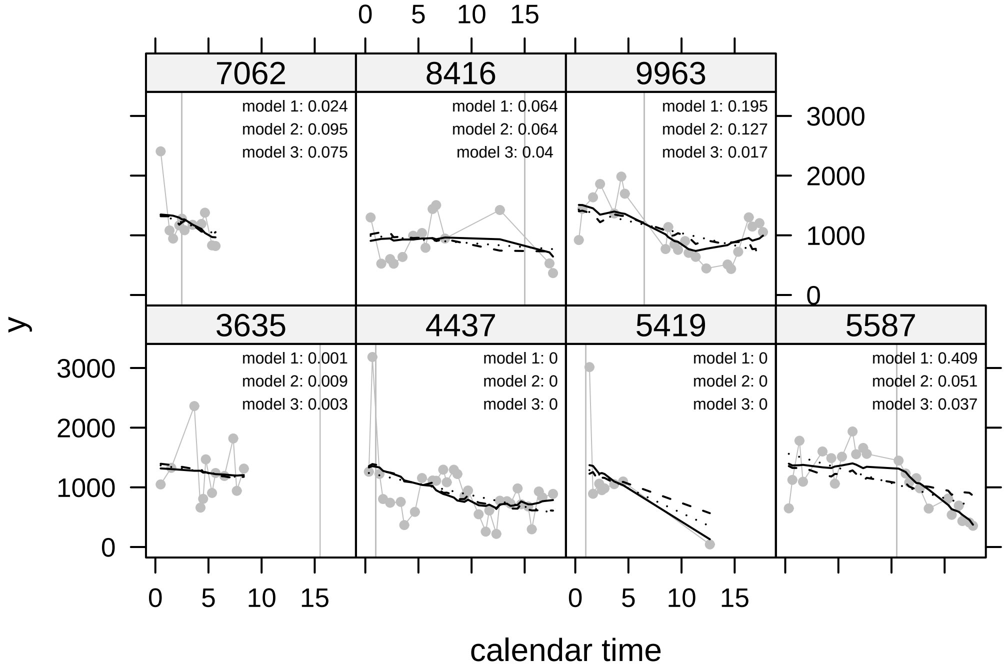


***Figure 4.14*** – *Model fits (estimated posterior means) for the patients with small posterior-predictive BOT values, which are noted in the top-right corners of the panels. The three models are discerned by line type: model 1 (——), model 2 (– – –), and model 3 (……).*

Another way to diagnose extreme observations is to look at the contributions of the observations to the mean posterior-predictive proper score of the model. For example, one could diagnose a multivariate trajectory as outlying if its score is outlying in the univariate sample of scores. Here, we instead examine the 6 observations with the highest scores in the three models, which correspond to the 5% worst scores each. Both for model 1 and for model 3, the 6 observations with the worst energy scores are IDs 3635, 4437, 5419, 5587, 8985 and 9963. For model 2, ID 8985 is replaced with ID 7146. These findings are similar to the BOT outliers: in model 1 and model 2, the IDs 3635, 4437 and 5419 were among those with BOT values less than 0.05. The 6 highest logarithmic scores are assigned to IDs 2376, 4437, 5587, 7288, 8418 and 8985 for model 2 and model 3. For model 1, ID 5587 is replaced with ID 7288.

We plot the trajectories of those IDs with high scores, which were not already plotted in Figure 4.14, in Figure 4.15. Strikingly all new "log-score outliers" are long time series ($n_i = 30, 23, 29, 25, 26, 27$), but their model fits do not look very strange. This shows that high log-scores alone are not indicative of outlying observations, because the absolute posterior-predictive density level is not indicative. Especially for our data set with observations of different dimensions $n_i$, the log-scores are not appropriate, because longer time series have a tendency to smaller density ordinates corresponding to higher log-scores. It is rather the *relative* density level of the materialized observation compared with the possible posterior-predictive density levels for the same individual, which is indicative of outliers – and this is exactly what the BOT values are. The posterior-predictive energy scores as generalized mean euclidean prediction errors are more appropriate than the log-scores. Their advantage over the BOT values is the easier Monte Carlo estimation and thus more general applicability, because they do not need known conditional densities.

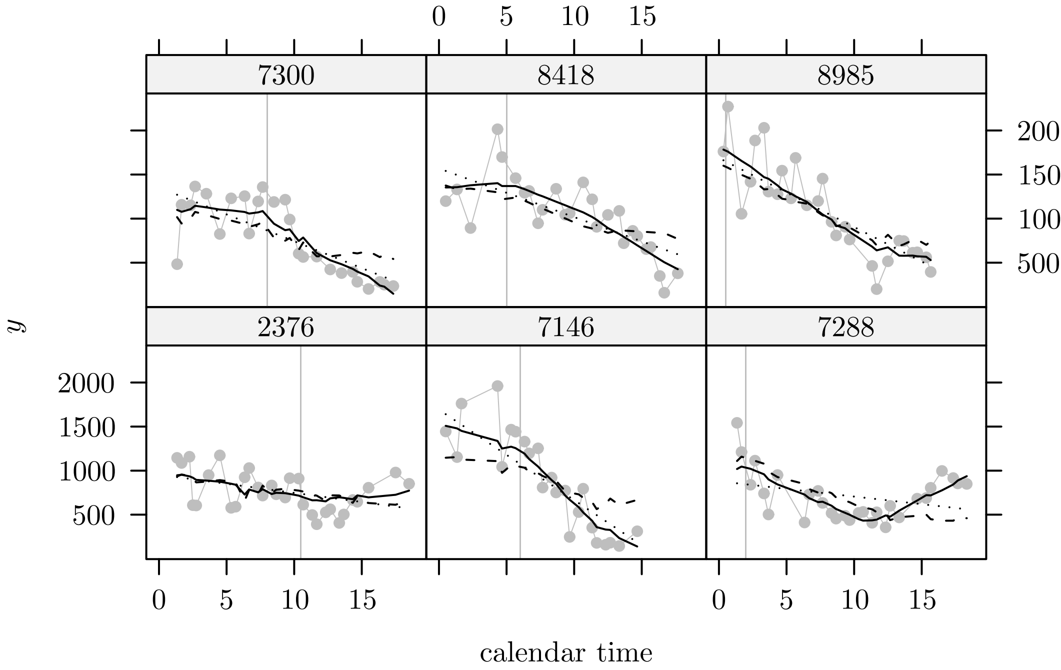


***Figure 4.15*** – *Model fits (estimated posterior means) for the patients with high posterior-predictive energy and logarithmic scores, which are not already plotted in Figure 4.14. The three models are discerned by line type: model 1 (——), model 2 (– – –), and model 3 (........).*

### Cross-validation assessment

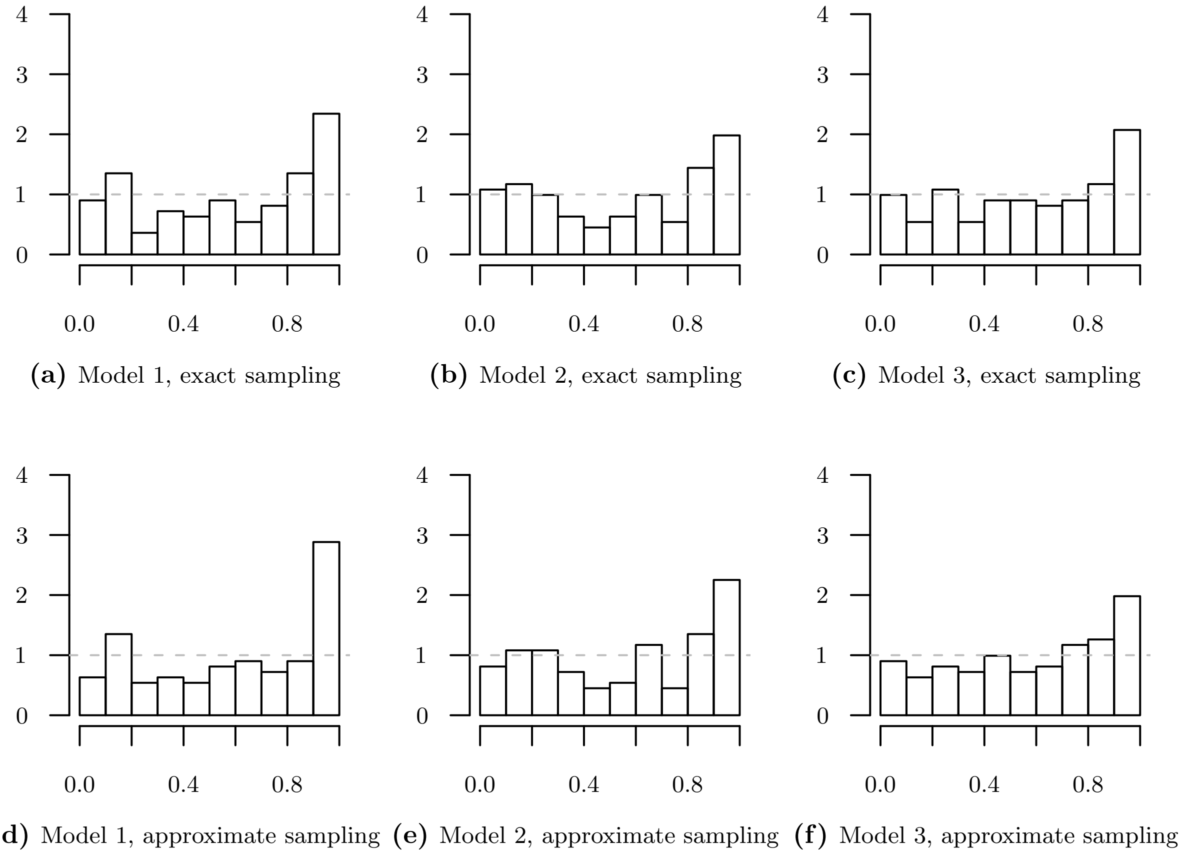
Next we want to check the leave-one-out predictive capabilities of the models. In order to check the performance of the approximate cross-validation scheme, we have also run the exact cross-validation scheme, and have saved every 20-th iteration out of chains of length 100 000 after a burn-in phase of 20 000 iterations. The approximate evaluations took 5, 2 and 3 seconds. The exact evaluations took much longer, with saved timings 14 643, 11 447 and 10 954 seconds for the three models. So the approximate sampling saves three orders of magnitude of computing time! However, it should be noted that a significant part of the required computing time for the exact cross-validation could be spared if the program would be integrated into `BayesX`, because the import overhead into `R` is memory-intensive. The situation is even worse on machines with small working memory when swap actions of the operating system are necessary. Nevertheless, the exact scheme will always be at least $n$ times slower than the approximate scheme, because the number of necessary Markov chains is $n$ instead of 1 (and this single chain is only necessary if the full model has not been sampled yet!).

In Figure 4.16 the BOT histograms from the exact and approximate sampling approaches are compared. All histograms show too large bars in the last bin $[0.9, 1.0]$, which means that the predictions for the left-out individuals are rather over- than underdispersed. Judging from the exact BOT histogram in panel (c), model 3 has the best calibration among the three models. The approximate counterpart in panel (f) is very similar. The exact BOT histograms for model 1 and model 2 in panels (a) and (b) are more left-skewed. This impression is even stronger in the approximate panels (d) and (e).

In Figure 4.17 we compare the energy and logarithmic scores resulting from the exact and approximate sampling schemes. It is surprising that the approximate logarithmic scores are almost perfectly matching the exact counterparts. For all three models, there is no noticeable departure from the identity line. The approximation of the energy scores seems to be slightly more difficult: especially for higher true scores and for model 3 in panel (c), the conservatism of the estimates is visible.

In Table 4.5 the mean scores are compared. Judging from the exact scores, model 2 is preferred over model 3 and model 1 by the mean energy score, while model 1 is preferred over model 3 and model 2 by the logarithmic scoring rule. These two rankings are reproduced by the approximate scores. The conservatism of the faster sampling scheme is conveyed by the absolute numbers: they are always smaller than the exact ones, with the relative error being larger for the energy scores. This behaviour is expected from the too optimistic nature of the approximate sampling strategy.

***Figure 4.16*** – *BOT histograms for calibration assessment of the leave-one-out prediction in the three random effects models. The predictive distributions were estimated with the exact (upper row, 4000 samples) and the approximate (lower row, 9000 samples) sampling schemes.*



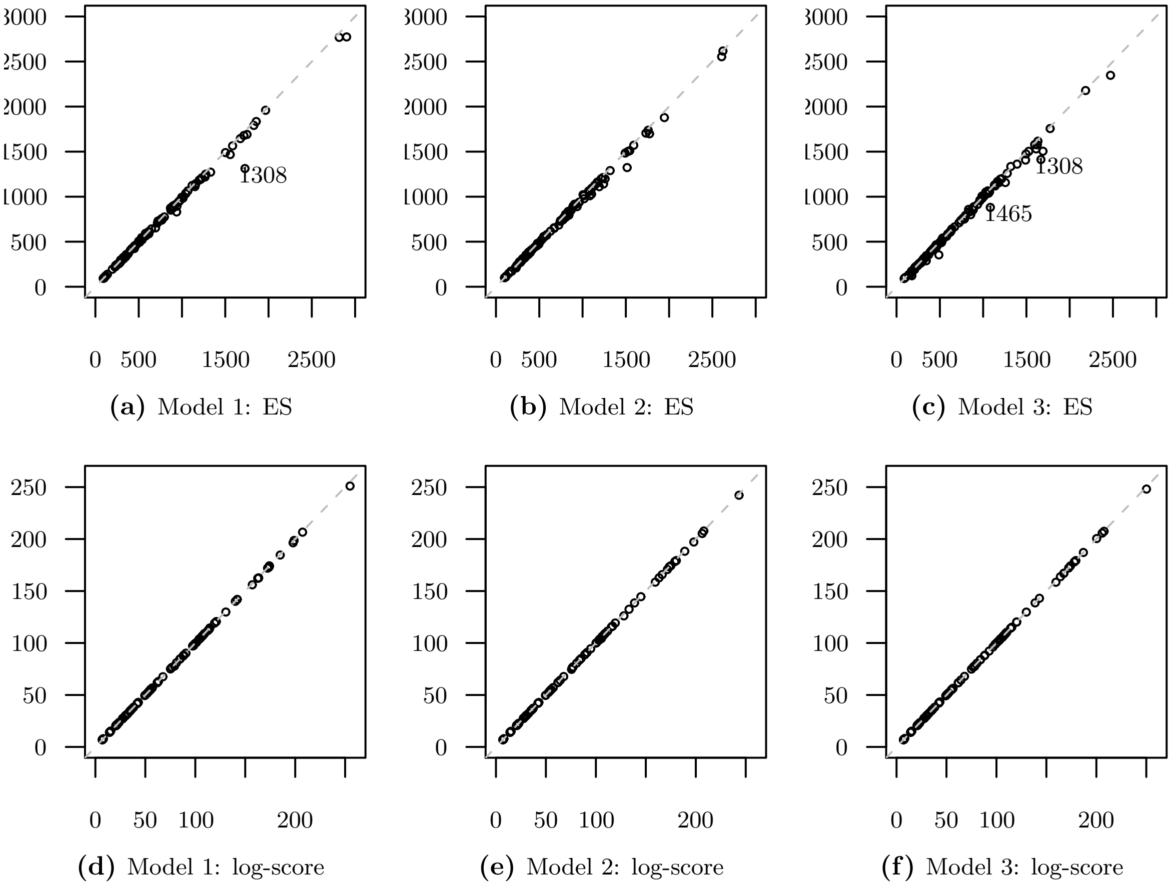
**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

***Table 4.5*** – *Mean energy and logarithmic scores for the cross-validated prediction of the three models, under the exact and approximate sampling schemes.*

| Scoring Rule | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| ES | exact | 758.76 | 733.65 | 748.41 |
|  | approximate | 740.36 | 717.03 | 728.95 |
| log-score | exact | 67.26 | 67.79 | 67.62 |
|  | approximate | 67.04 | 67.58 | 67.52 |

***Figure 4.17*** – *Comparison of exact and approximate scores for leave-one-out prediction in the three random effects models (columns). The panels in the upper row compare the energy scores (ES), while the panels in the lower row compare the log-scores. Individuals where the absolute difference between the exact and approximate energy score values exceeds 200 are labelled.*



**(a)** Model 1: ES **(b)** Model 2: ES **(c)** Model 3: ES

**(d)** Model 1: log-score **(e)** Model 2: log-score **(f)** Model 3: log-score

### Results

While the goodness-of-fit histograms did not favour model 2, its DIC and mean energy score are the best of all three models. The calibration of the leave-one-out predictions also appeared acceptable. This is another example for the well-known fact that the best-fitting model for the known data is not necessarily the best-predicting model for new data. However, the decision is not totally straightforward here because the mean logarithmic score actually ranks model 2 worst. For a description of the known data, model 1 might be better suited, because it has a better fit to the known data than model 2 and the best leave-one-out log-score.

We show the posterior summaries for the fixed effects from model 2 in Table 4.6. The psychological score (PSSCO) has a significant association with the CD4 cell counts according to the table, with a higher score representing negative feelings being associated with a lower cell count. Cigarette smoking during the last six months before the interview does not seem to have an equally intuitive association with the response, as the Bayesian point estimates of the dummy variable coefficients (PACKS...) are positive (which *could* theoretically result from a causal "smoking increases CD4 counts" relation) but the posterior distributions are centered around zero. The number of sexual partners (NSEX) seems to have a strong positive association with the dependent variable, while the modelled effect of recreational drugs usage (DRUGSyes) is not statistically significant, because a positive and negative sign for the coefficient are almost equally probable *a posteriori* (59% vs. 41%).

| Coefficient | Mean | Median | SD | lower | upper | Positive |
|---|---|---|---|---|---|---|
| TIME | −21.19 | −21.16 | 4.25 | −29.35 | −12.72 | 0.00 |
| TIMEpos | −45.80 | −45.86 | 6.80 | −58.61 | −31.99 | 0.00 |
| TIMEposLate | 57.27 | 57.32 | 10.24 | 37.86 | 77.77 | 1.00 |
| PSSCO | −83.72 | −84.12 | 25.10 | −132.14 | −34.41 | 0.00 |
| PACKShalf | 40.73 | 40.47 | 167.72 | −287.23 | 366.97 | 0.60 |
| PACKSone | 37.41 | 38.71 | 166.28 | −295.03 | 351.14 | 0.59 |
| PACKStwo | 97.97 | 97.79 | 165.77 | −226.94 | 418.15 | 0.72 |
| PACKSmore | 44.05 | 44.44 | 170.18 | −290.89 | 375.77 | 0.60 |
| NSEXone | 67.62 | 68.12 | 60.13 | −53.16 | 181.85 | 0.87 |
| NSEXmore | 105.26 | 105.36 | 59.34 | −13.24 | 219.12 | 0.96 |
| DRUGSyes | 30.62 | 31.21 | 130.94 | −215.93 | 295.68 | 0.59 |

***Table 4.6*** – *Posterior summaries for fixed effects coefficients in model 2: In addition to the posterior mean, median and standard deviation of the coefficient, the lower and upper bound of the 95% HPD-interval and the posterior probability that the coefficient is positive are shown.*

The proposed approximate sampling scheme yielded very good results in this data example: the logarithmic scores were approximated very well, and the energy scores approximations were only slightly worse. The leave-one-out BOT histograms were more difficult to approximate, but the general calibration picture was retained under the parsimonious sampling scheme.

### 4.5.2 Analysis for all patients with covariate time

Given the promising results on the performance of the approximate cross-validation scheme from the last section, we now want to analyze the data from all patients with respect to the association of the fully available covariate time with the CD4 cells counts.

#### Model fitting

We will sample from the posterior distribution in six selected models, which are listed in Table 4.7. Model 2 differs from model 1 in that it assumes a linear time effect only after the seroconversion date. The idea is that the CD4 counts are constant before the HIV infection. This assumption is also coded into model 3 and model 4, which both feature a second basis function taking effect in seroconverters from 1995 on. The variables `TIMEpos2` and `TIMEposLate2` are just the squares of the linear bases `TIMEpos` and `TIMEposLate`. The resulting time trends for model 4 are continuous. More flexible fixed time trends are allowed in model 5 and model 6, where P-splines (Brezger and Lang 2006) are used. Model 6 adds linear random effects as in model 4. Note that the option `nofixed` is used to disable the incorporation of analogous fixed effects. We do not want them because we already have the flexible P-spline modelling the fixed time effect, and adding another base could lead to Markov chain convergence difficulties due to weakly identified parameters.

| No. | `BayesX` predictor formula |
|---|---|
| 1 | `CASEID`$(random)$ + `TIME` ∗ `CASEID`$(random)$ |
| 2 | `CASEID`$(random)$ + `TIMEpos` ∗ `CASEID`$(random)$ |
| 3 | `CASEID`$(random)$ + `TIMEpos` ∗ `CASEID`$(random)$ + `TIMEposLate` ∗ `CASEID`$(random)$ |
| 4 | `CASEID`$(random)$ + `TIMEpos2` ∗ `CASEID`$(random)$ + `TIMEposLate2` ∗ `CASEID`$(random)$ |
| 5 | `TIME`$(psplinerw2, nrknots = 5)$ + `CASEID`$(random)$ |
| 6 | `TIME`$(psplinerw2, nrknots = 8)$ + `CASEID`$(random)$ + `TIMEpos` ∗ `CASEID`$(random, nofixed)$ + `TIMEposLate` ∗ `CASEID`$(random, nofixed)$ |

***Table 4.7*** – *Overview of the six `BayesX` models for the response variable `LEU3N`.*

We produced Markov chains of length 200 000 and saved every 20-th iteration for all six models, but only after the burn-in phase of 100 000 iterations. We discarded the burn-in directly in `BayesX` to reduce the memory allocation load for the import into `R`, which is quite high due to the large number ($\sum_{i=1}^{n} n_i = 10\,606$) of data points. Traceplots and cumulative quantile plots were checked to ensure that the used burn-in was large enough.

We plot the estimated time trends in Figure 4.18. Note that the pointwise and simultaneous credible intervals are almost identical here. The P-splines in panel (e) for model 5 and in panel (f) for model 6 feature the typical curve form with an inflexion point around 2.5 years after seroconversion, and curved to the right and left before and after seroconversion, respectively. This form is mimicked in panel (c) by the TP spline with knots at seroconversion and 1995, and constant level before seroconversion. However, the fixed effect is invariant to the calendar time in model 5 and model 6, so strictly the trends are not directly comparable to the trend in model 3. Model 4 in panel (d) shows a problem of the model, as it fits negative time effects near the end of the exemplary time scale. As there are no other covariates (but random time effects) in the models, this corresponds to negative mean CD4 counts. An alternative would thus be to logarithmize the CD4 counts and model them instead of the original counts, which is discussed in section 4.7. The trends in model 1 and model 2 shown in panels (a) and (b) are rather too simple compared with the P-spline trends.

### Goodness-of-fit assessment

In order to assess the goodness-of-fits, we look at posterior-predictive PIT histograms in Figure 4.19. In all plots, the last bar for the bin $(0.95, 1]$ is remarkably larger than the bars to its left. This indicates that some individual observations are clearly underestimated by the models. Model 3 in panel (c) and model 6 in panel (f) have more PIT values near 0.5 than the other models. However the differences between the histograms are small.

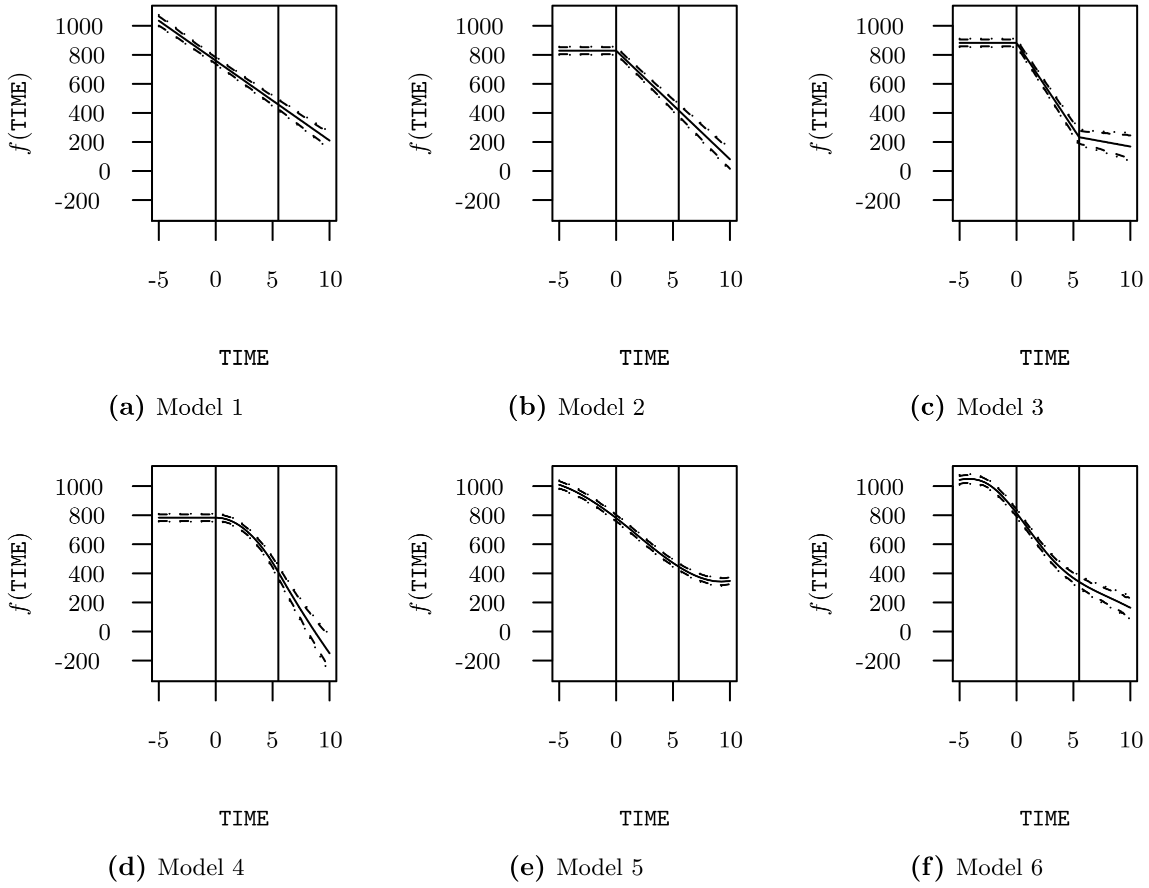
In Table 4.8 the posterior-predictive mean scores are listed. The scores support the PIT histograms, because model 6 has the lowest scores and is thus ranked as the model with the best fit by the posterior-predictive scores. The second-best fit is provided by model 3. It is interesting that for this data, the energy and the logarithmic score agree on the goodness-of-fit ranking of all models.

***Table 4.8*** – *Posterior-predictive mean energy and logarithmic scores for the goodness-of-fit assessment of the six models.*

| Fit criterion | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| ES | 587.26 | 604.48 | 541.33 | 629.37 | 673.29 | 531.91 |
| Log-score | 125.90 | 126.45 | 124.22 | 126.91 | 128.18 | 123.84 |

We show the 15 observations having a posterior-predictive BOT value of zero in all of the six models in Figure 4.20. Only two IDs (4437 and 5419) were already included in the outlier Figure 4.14 from the complete data analysis, where other covariates were considered. Individual time series with large jumps in the CD4 counts, and long time series with clear non-linearity are obviously most difficult to fit, for all six considered

***Figure 4.18*** – *Estimated fixed effects time trends in the six models: Means (___), pointwise HPD (- - -), and simultaneous (.......) credible intervals at the 0.95 level are plotted. The vertical lines mark the seroconversion date (mid-1989) and the year 1995 for the virtual average patient (random effects are not included) having entered the study in mid-1984.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

**(d)** Model 4 **(e)** Model 5 **(f)** Model 6

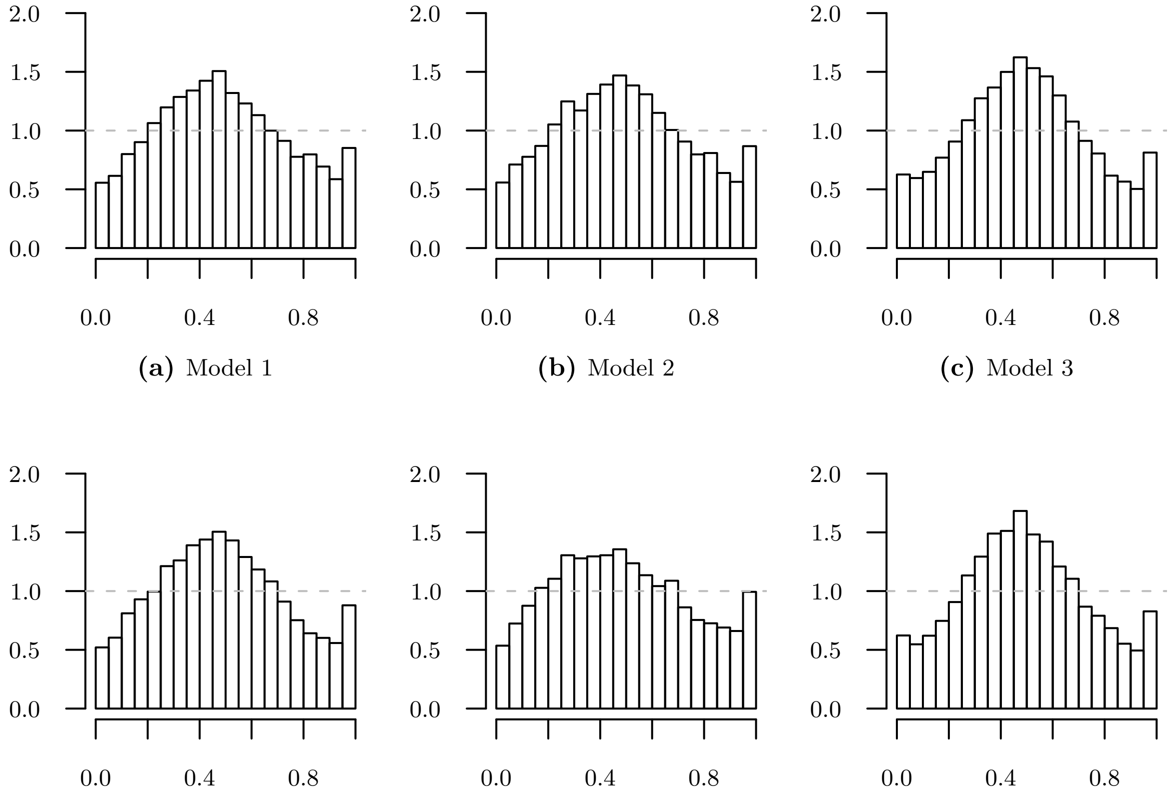
models.

## Cross-validation assessment

An exact leave-one-out cross-validation assessment for each of the six models is infeasible: For example, because already the model fit of model 1 took 1253 seconds, the iterative fit of all 574 reduced models with the model 1 predictor form would take approximately 719 140 seconds or 200 hours. Instead, we trust the approximate sampling strategy, and generate according samples for the six models in merely 8.1, 7.1, 8.5, 8.5, 5.8 and 8.6 seconds, respectively.

We check the calibration with the BOT histograms in Figure 4.21. All six histograms are

***Figure 4.19** – Scalar PIT histograms for goodness-of-fit assessment of the six random effects models*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

**(d)** Model 4 **(e)** Model 5 **(f)** Model 6

far away from an optimal uniform distribution of the BOT values. Only large fractions for the lowest bin $[0, 0.1]$ indicate that the histograms do not show the distribution of posterior-predictive, but leave-one-out BOT values.

Marshall and Spiegelhalter (2007, p. 429) recommend to also check for outliers using the approximate leave-one-out samples, which they call “mixed predictive samples”, if one is concerned with the random effects prior distribution. They use the posterior-predictive samples to check the likelihood assumptions, which comprises the form of the linear predictor or equivalently the model in our application. (Of course also the identity link and the normal distribution assumption are part of the likelihood assumptions.) If we look for individuals having a cross-validation BOT value of 0, we find that the IDs are 2545, 2755, 3861, 7398, 8269, 8854 and 9888. This means that 6 of the 7 individuals were already included in the posterior-predictive BOT outliers.

Turning to the model choice, `BayesX` reports DIC values. We are interested if the

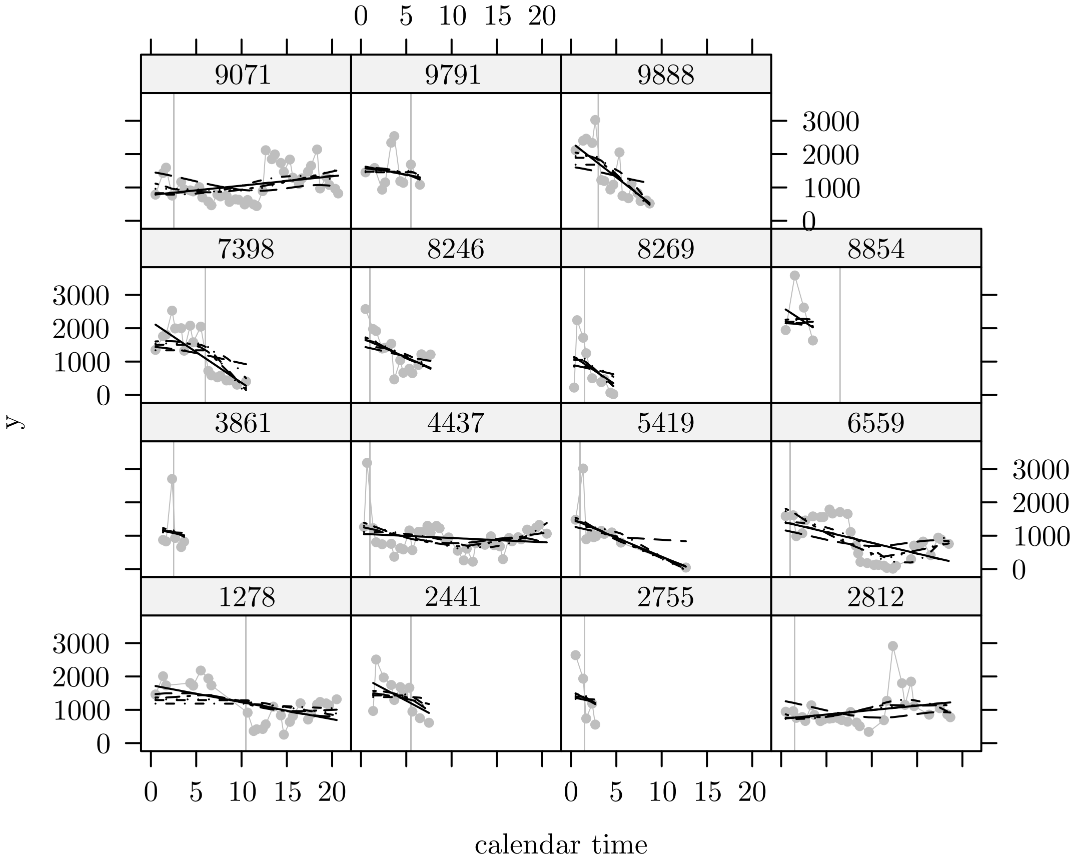
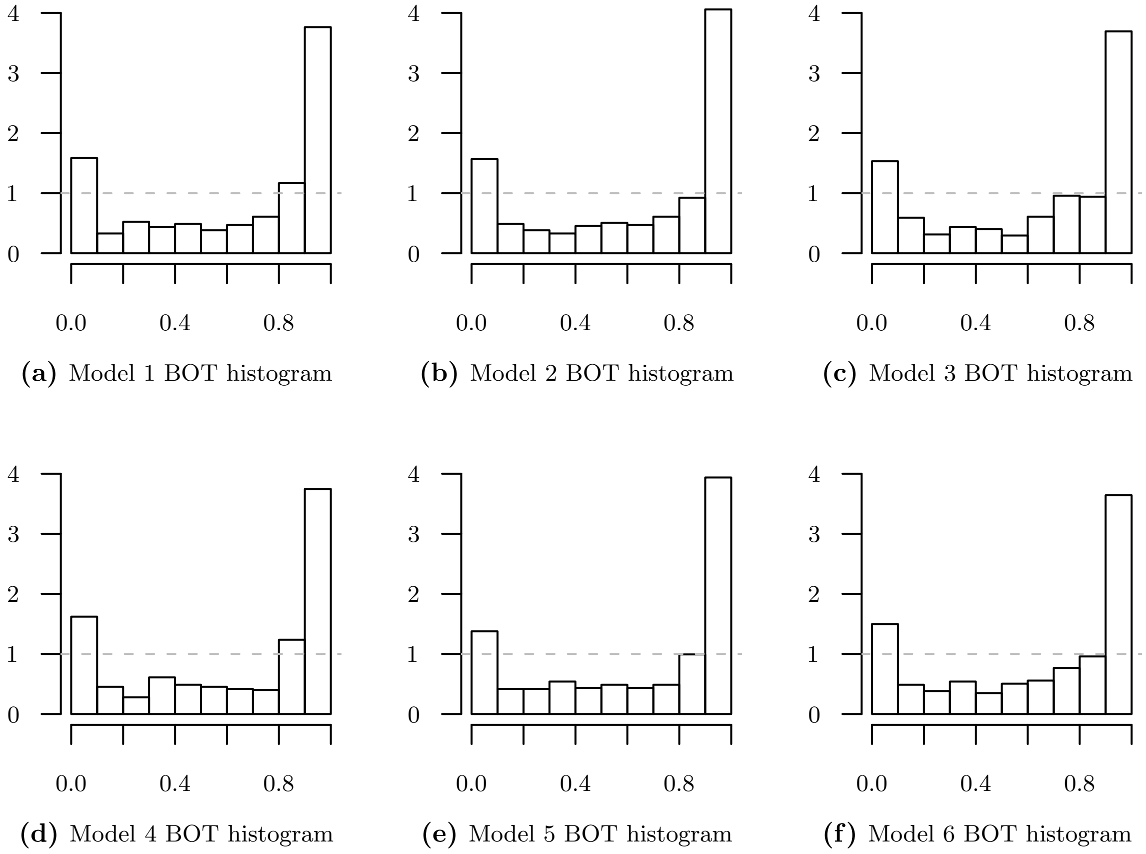


***Figure 4.20*** – *Model fits (estimated posterior means) for the patients with posterior-predictive BOT values of zero. The six models are discerned by line type: model 1 (——), model 2 (- - -), model 3 (........), model 4 (._._), model 5 (----), and model 6 (.—.—).*

approximate leave-one-out proper scoring rules yield the same result. In Table 4.9 the mean energy and logarithmic scores of the models in question as well as the DIC values are shown. Model 5 is preferred by the DIC (followed by models 2, 1, 4, 6, 3), and also by the energy score (followed by models 6, 1, 3, 2, 4). By contrast, the logarithmic scoring rule ranks model 6 best (followed by models 3, 1, 2, 5, 4). This is a large difference to the DIC ranking, e. g. model 6 is up from the last-but-one place and model 5 is down from the first place to the last-but-one place. Yet, all three criteria agree that a P-spline model should be chosen.

*Figure 4.21* – *Approximate BOT histograms for calibration checking of the leave-one-out predictions in the six models.*



(a) Model 1 BOT histogram

(b) Model 2 BOT histogram

(c) Model 3 BOT histogram

(d) Model 4 BOT histogram

(e) Model 5 BOT histogram

(f) Model 6 BOT histogram

*Table 4.9* – *Approximate mean energy and logarithmic scores for the cross-validated prediction of the six models, as well as the DIC based on the saturated deviance samples reported by* `BayesX`.

| Model criterion | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| ES | 1067.11 | 1133.65 | 1108.70 | 1539.25 | 904.44 | 1015.88 |
| Log-score | 128.87 | 129.41 | 127.82 | 130.34 | 129.65 | 127.25 |
| DIC | 11555.5 | 11531.8 | 11694.0 | 11621.1 | 11137.0 | 11669.2 |

### Results

The best fit to the known data is provided by model 6. As this model does also perform well in the approximate leave-one-out cross-validation (under both proper scoring rules), we would probably choose model 6 for the description of the known data. Model 5 does not have an equally good fit, but could be an alternative, because the DIC and the energy score prefer it over model 6.

## 4.5.3 Final model

The final model for the CD4 data is a synthesis of model 6 from section 4.5.2 and model 2 from section 4.5.1: we include the significant covariates PSSCO and NSEX (cf. Table 4.6) into the P-spline model. We omit the covariates PACKS and DRUGS to keep the resulting data set as large as possible. This final model thus has the BayesX formula

$$\begin{aligned}\texttt{LEU3N} = {} & \texttt{PSSCO} + \texttt{NSEXone} + \texttt{NSEXmore} + \texttt{CASEID}(random) + \\ & \texttt{TIME}(psplinerw2, nrknots = 8) + \texttt{TIMEpos} * \texttt{CASEID}(random, nofixed) + \\ & \texttt{TIMEposLate} * \texttt{CASEID}(random, nofixed).\end{aligned}$$

The reduced data set comprises all $n = 574$ participants, but only $\sum n_i = 6478$ data points (minimum 1, maximum 36 observations per participant). We produced a total of 5000 parameter samples for this model specification by thinning out a Markov chain of length 200 000 and discarding a burn-in of 100 000 iterations, within 1125 seconds.

The posterior summaries of the fixed effects are tabulated in Table 4.10. Compared with Table 4.6 on page 128 for the fixed effects in model 2 from the previous section, the direction of the estimated associations is unaltered: Worse psychological scores are significantly associated with decreasing CD4 cell counts, and the number of sexual partners in the last six months is positively correlated with the CD4 cell counts. Note that the posterior means of the coefficients are different because the covariates PACKS and DRUGS as well as the fixed parametric effects for TIME have been omitted, and instead a fixed P-spline effect has entered the model.

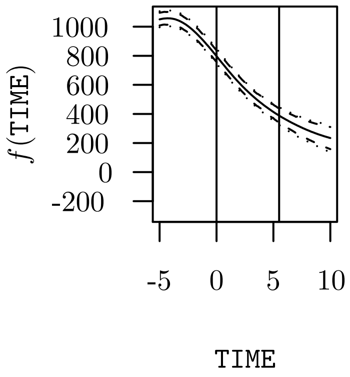
The fixed effect time trend estimate from that P-spline is graphed in Figure 4.22. The trend is very similar to the model 6 trend in panel (f) on page 131, but the credible intervals are wider. This is comprehensible, because the time-varying covariates PSSCO and NSEX have entered the model, and more model parameters lead to larger uncertainty about the covariates' associations.

Next we want to check the goodness-of-fit of the new model. The posterior-predictive BOT and PIT histograms are shown in Figure 4.23. If we compare the PIT histogram in

***Table 4.10*** – *Posterior summaries for fixed effects coefficients in the final model: In addition to the posterior mean, median and standard deviation of the coefficient, the lower and upper bound of the 95% HPD-interval and the posterior probability that the coefficient is positive are shown.*

| Coefficient | Mean | Median | SD | lower | upper | Positive |
|---|---|---|---|---|---|---|
| PSSCO | −46.37 | −46.30 | 9.01 | −63.01 | −27.91 | 0.00 |
| NSEXone | 22.40 | 22.34 | 17.37 | −11.05 | 55.33 | 0.90 |
| NSEXmore | 34.21 | 34.42 | 17.40 | 0.04 | 67.87 | 0.97 |

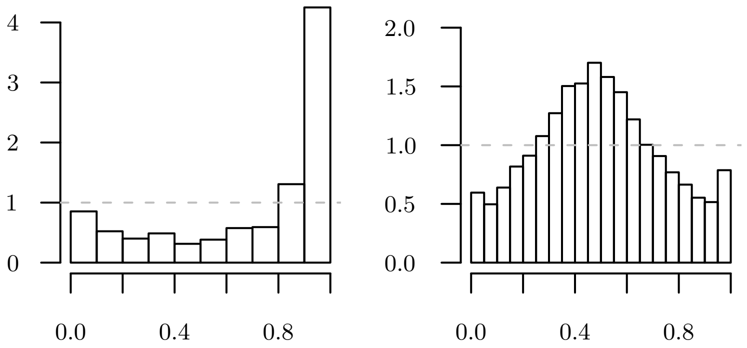
***Figure 4.22*** – *Estimated fixed effects time trend in the final model: Means (——), pointwise HPD (- - -), and simultaneous (........) credible intervals at the 0.95 level are plotted. The vertical lines mark the seroconversion date (mid-1989) and the year 1995 for the virtual patient (random effects are not included) having entered the study in mid-1984, who has constant covariate values* NSEX == *none and* PSSCO == 1.



panel (b) with the model 6 PIT histogram in panel (f) on page 132, we do not observe large differences. The BOT histogram in panel (a) shows a high frequency of large BOT values, and thus also shows a good fit of the new model to the given data.

In order to check the leave-one-out calibration of the final model, we plot the approximate cross-validation BOT and PIT histograms in Figure 4.24, as the exact cross-validation would have required ca. 179 hours. The BOT histogram in panel (a) shows a better calibration than the BOT histogram in panel (f) on page 134 for the model with covariate time only. Note however that while we have all individuals in the data set here, there are fewer data points attached to the individuals. So the sample leading to the BOT histogram has the same size, but it is smaller for the PIT histogram in panel (b). It shows a relatively good calibration for the scalar predictive distributions, which do seem to have too heavy lower tails compared to the materialized observations. This is indicated by the small bars for the lower bins: too few observations materialize in the lower tails.

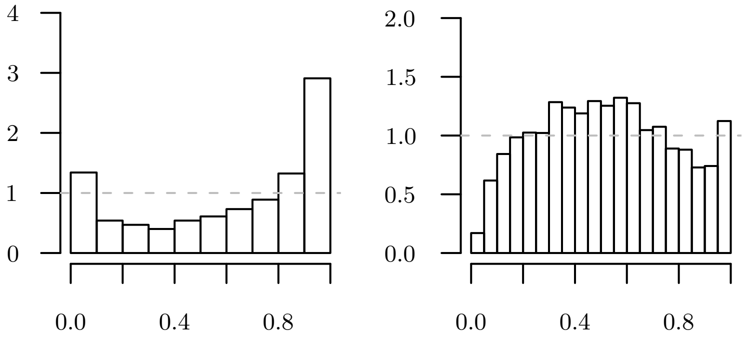
***Figure 4.23*** – *Posterior-predictive BOT and PIT histograms for goodness-of-fit assessment of the final model.*



**(a)** BOT histogram

**(b)** PIT histogram

***Figure 4.24*** – *Approximate BOT and PIT histograms for leave-one-out predictive calibration assessment of the final model.*



**(a)** BOT histogram

**(b)** PIT histogram

## 4.6 BMI data

The BMI data set is a subset of the LISA[vi] study data, a recent German birth cohort study originally designed to examine influence factors for the development of the immune system and allergies in children (Jacob et al. 1999). We follow Fenske, Fahrmeir, Rzehak, and Höhle (2008) and instead investigate the $n = 2043$ children's body mass index (BMI) dependence on risk factors already discussed in the literature. The corresponding covariates are listed in Table 4.11 and are time-constant, except for the time variable `age`. There are between $\min n_i = 2$ and $\max n_i = 9$ individual observations for each child, between its birth and an age of $\max t_{ij} = 6.31$ years. This gives the total number of data points $\sum n_i = 17\,316$. More details on the data set are given by Fenske, Fahrmeir, Rzehak, and Höhle (2008, p. 3).

| Covariate | Description |
|---|---|
| `age` | Age in years |
| `wgain2y` | Weight gain until the age of 2 years |
| `tvpc` | Hours spent watching TV and playing computer at the age of 4 years (4 classes) |
| `outdoor` | Hours spent outdoor per day at the age of 4 years |
| `mEdu` | Maternal highest level of education (5 classes) |
| `mBMI` | Maternal BMI at pregnancy begin |
| `mDiffBMI` | Maternal BMI gain during pregnancy |
| `mSmoke` | Did the mother smoke during pregnancy? |
| `breast` | Bottle-feed and/or breastfeeding, or breastfeeding only? |
| `area` | Rural or urban study centre? |

***Table 4.11*** *– Description of BMI data set covariates.*

In Figure 4.25 we show the trajectories of 18 randomly selected children. Typically the BMI levels rise until the age of around 1 year and decline slowly afterwards. However, there are also children whose BMI is highest at the end of the study time, e. g. IDs 92182314 and 95084041. Note that the absolute calendar time is not relevant for this data set, because the neonates were recruited within in the short time of 15 months, and can thus be treated as a time-homogeneous cohort.

[vi] LISA is the abbreviation of "Influences of *L*ife-style factors on the development of the *I*mmune *S*ystem and *A*llergies in East and West Germany"

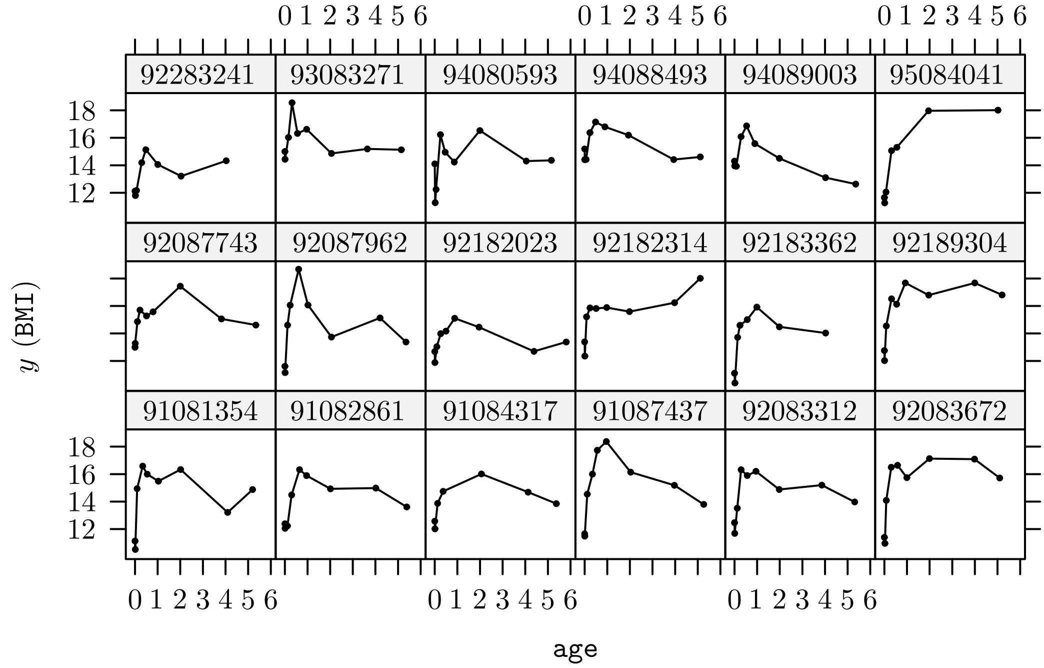


*Figure 4.25* – *BMI trajectories for a random subset of 18 children, whose IDs are in the headings of the panels.*

### 4.6.1 Approximate sampling performance case study

First, we will check the performance of the approximate sampling scheme. To this end, we take the first 100 children from the complete data set. The number of observations per child ranges between $\min n_i = 4$ and $\max n_i = 9$, giving a total of $\sum n_i = 829$ data points in the subsample. The restriction to this subsample is necessary to do the exact leave-one-out cross-validation in a manageable amount of time. See section 4.6.2 for an analysis of the whole data set.

#### Model fitting

For all three candidate models, we include dummy variables for the binary covariates `sex` and `mSmoke`. Moreover, we include a binary variable `mEduHigh` which is 1 if the mother has Abitur or Fachabitur and is 0 else, that is we collate the two highest education levels and contrast them with the lower three levels of `mEdu`. The variable `tvpcMoreThan1` analogously collates the highest three levels of the ordinal variable `tvpc`. The continuous variables `mBMI` and `outdoor` are included as well. These covariate choices are of course quite arbitrary, but are made to simplify the case study. All available covariates will later be used in section 4.6.2.

The three models feature a random intercept, and differ in the modelling $f(\texttt{age})$ of the time variable `age`: For the first model, we assume a hockey-stick form $f(t) := \alpha \cdot t + \beta \cdot (t-1)_+$ with the breakpoint at the age of one year ($t = 1$). Both basis functions (which are called `age` and `ageAfter1`) get fixed and random effects (parameters $\alpha$, $\beta$ and $\alpha_i$, $\beta_i$, $i = 1, \dots, n$, say), to adjust for unexplained heterogeneity between the children. For the second model, a more sophisticated parametric form $f(\texttt{age})$ for the `age` variable is used, which is inspired by the typical trajectories we have already seen in Figure 4.25:

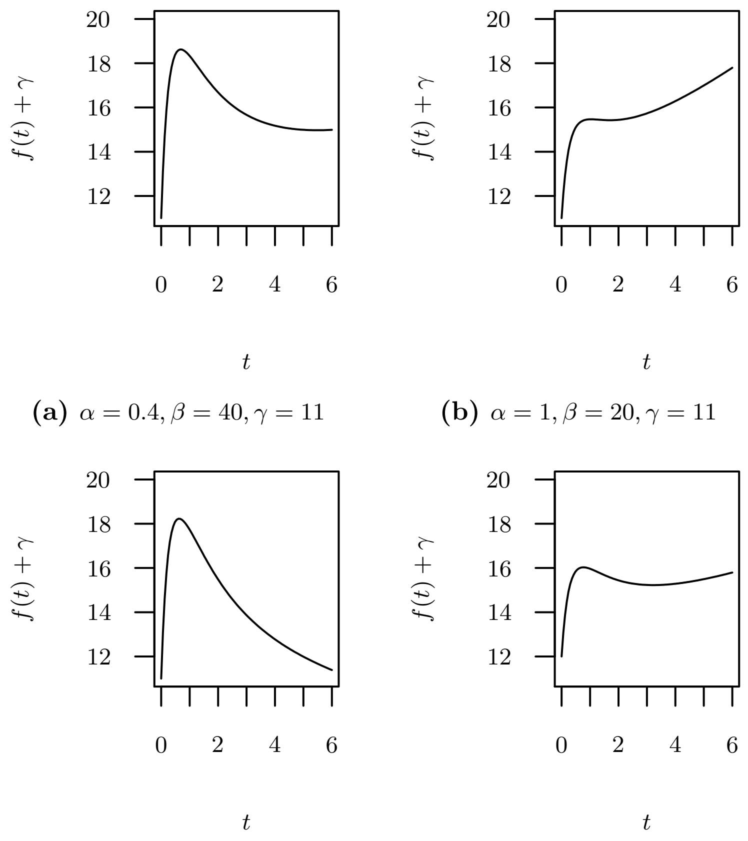
$$f(t) := \alpha \cdot t + \beta \cdot \frac{\log(t+1)}{(t+1)^2} \tag{4.6.1}$$

Together with the intercept, say $\gamma$, this function can approximate many typical features, as Figure 4.26 shows. Both coefficients $\alpha$ and $\beta$ are included as fixed and random effects. The third model uses a P-spline with 8 knots to model the fixed time effect. In addition, a random slope is included in the model.

For each of the three models, we ran `BayesX` with a burn-in phase of 100 000 iterations, after which every 20-th sample of the next 200 000 iterations was saved. The overall convergence was successfully checked with deviance traceplots. For model 1, strong autocorrelations between the fixed effect samples of the `age` and `ageAfter1` covariates could be diagnosed. Since we intend to make posterior inference only for the whole time predictor function, but not for these single coefficients, this should not concern us unduly.

We plot the estimated time trends in Figure 4.27. The P-spline fit from model 3 in panel (c) looks too wiggly, compared with the two other parametric fits. The trend could be smoothed stronger, if we specified other hyperparameters for the P-spline variance prior: We used the default parameters $a = b = 0.001$ for the inverse-gamma prior, giving a prior mode of $b/(a+1) \approx 0.001$. For example, setting these values to $a = b = 0.0001$ decreases the mode to 0.0001, leading to a stronger penalization of second-order differences of the B-spline basis functions coefficients. Alternatively, the number of knots could be set lower than 8. The parametric fit from model 2 in panel (b) is similar to the P-spline fit for the age under one year. Afterwards, it is much smoother, which is of course implied by the strong parametric assumptions of the form (4.6.1). The linear TP-spline fit from model 1 in panel (a) is even more simple than the fit from model 2, but is still better interpretable than the model 3 fit: sharp increase of BMI until the age of one year, and slow decrease afterwards.

***Figure 4.26*** – *Possible parametric functions of the form* (4.6.1) *in model 2, when the intercept $\gamma$ is included.*



**(a)** $\alpha = 0.4, \beta = 40, \gamma = 11$

**(b)** $\alpha = 1, \beta = 20, \gamma = 11$

**(c)** $\alpha = -0.2, \beta = 40, \gamma = 11$

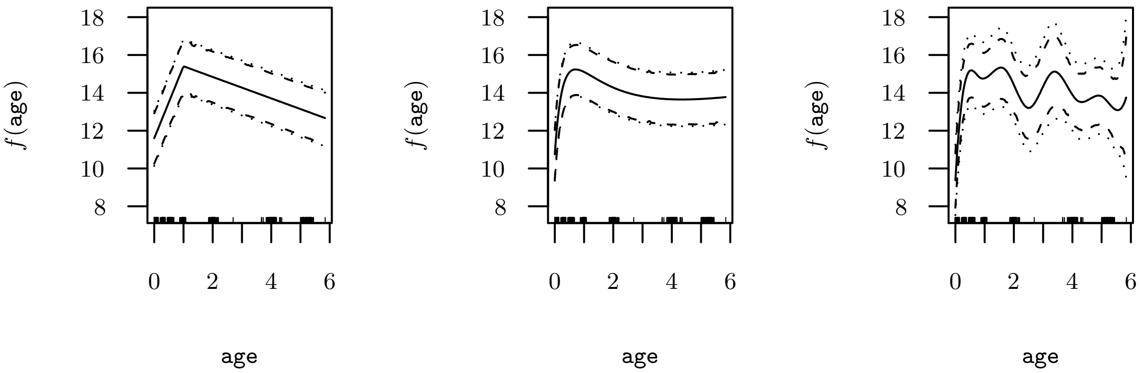
**(d)** $\alpha = 0.5, \beta = 20, \gamma = 12$

### Goodness-of-fit assessment

It is interesting what the different time trends can contribute to fitting the given data. The goodness-of-fit for the individual scalar observations and the whole vector-valued time series can be assessed with the posterior-predictive PIT and BOT histograms in Figure 4.28. The model 2 PIT histogram in panel (e) looks best among the three models: For model 3 in panel (f), more posterior-predictive PIT values below 0.2 have been observed, and for model 1 in panel (d), we have rather a uniform than a hump-shaped histogram. The BOT histograms convey the same statement. The model 2 histogram in panel (b) is stronger left-skewed than the model 3 and model 1 histograms in panels (c) and (a), respectively. So model 2 fits the given data best, followed by model 3 and model 1, if we measure the goodness-of-fit with the scalar observations (PIT) or on the predictive multivariate density scale (BOT).

In Table 4.12 the posterior-predictive mean scores are listed. The numbers match the

***Figure 4.27*** – *Estimated fixed effects time trends (including the intercept) in the three models: Means (——), pointwise HPD (- - -), and simultaneous (……) credible intervals at the 0.95 level are plotted. The positions of the x-coordinates are included in the form of x-axis ticks.*



**(a)** Model 1 **(b)** Model 2 **(c)** Model 3

impression from the PIT and BOT histograms: the goodness-of-fit is best for model 2, followed by model 3 and then model 1.

***Table 4.12*** – *Posterior-predictive mean energy and logarithmic scores for the goodness-of-fit assessment of the three models for the BMI subsample data.*

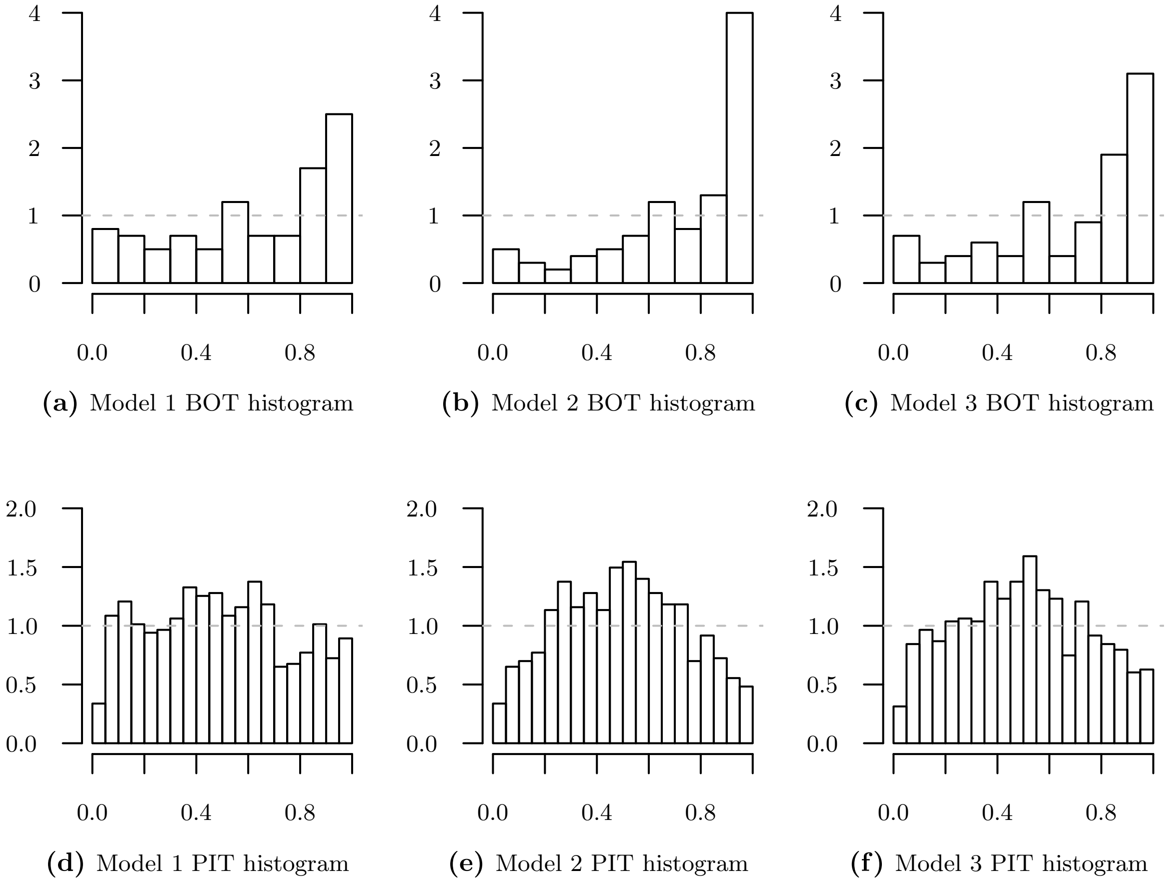
| Fit criterion | Model 1 | Model 2 | Model 3 |
|---|---:|---:|---:|
| ES | 2.75 | 1.72 | 1.86 |
| Log-score | 14.81 | 11.23 | 11.75 |

### Cross-validation assessment

We produced both approximate and exact leave-one-out model parameter samples for a predictive assessment of the three models. While the exact cross-validation sampling was very computer-intensive (9281, 9296 and 9759 seconds), the approximate sampling based on the samples from the model fitting was quickly done (4, 3 and 3 seconds). For the exact sampling, for each left out child, we used `BayesX` to produce chains of length 100 000, which were thinned out with parameter 20 after a burn-in phase of 20 000 iterations. No convergence problems were found in randomly selected traceplots of the resulting means and precisions samples.

The BOT histograms from both sampling schemes are compared in Figure 4.29. The exact BOT histograms show that all three models are quite well calibrated. The differences

***Figure 4.28*** – *Multivariate BOT (upper row) and scalar PIT (lower row) histograms for goodness-of-fit assessment of the three random effects models (columns) for the BMI subsample data.*



**(a)** Model 1 BOT histogram

**(b)** Model 2 BOT histogram

**(c)** Model 3 BOT histogram

**(d)** Model 1 PIT histogram

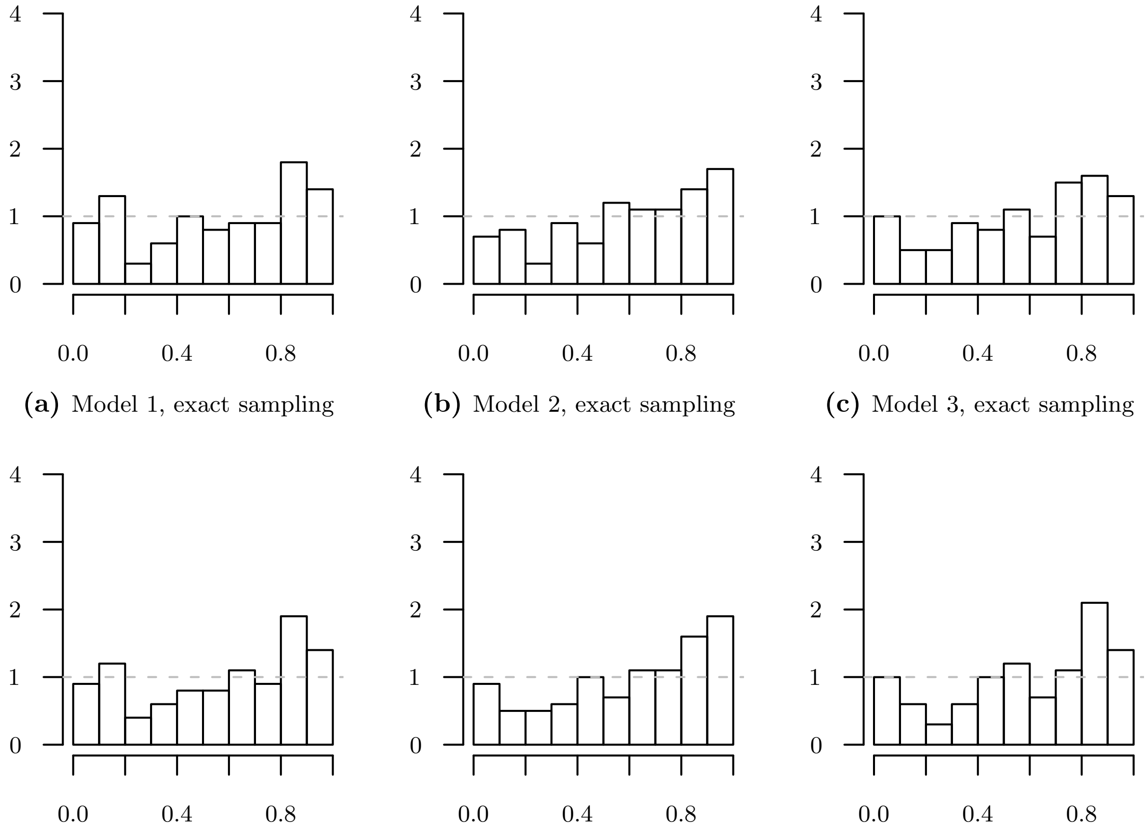
**(e)** Model 2 PIT histogram

**(f)** Model 3 PIT histogram

are rather small, with the histograms for model 1 in panel (a) and model 2 in panel (b) looking slightly better than the histogram for model 3 in panel (c). For model 1 and model 2, the approximate histograms in panels (d) and (e) are very near to the exact counterparts. The model 3 approximate BOT histogram in panel (f) is even a bit more left-skewed than the exact histogram in panel (c), so the model calibration ranking would be the same if we only had available the approximate BOT histograms.

In Figure 4.30 we compare the energy and logarithmic scores resulting from the exact and approximate sampling schemes. The logarithmic scores approximation works, although the (relative) deviances from the identity lines are larger than in Figure 4.17 for the CD4 data example. The logarithmic scores are approximated best in model 1, as panel (d) shows. The comparison plots for the energy scores are also slightly worse than the plots in Figure 4.17. Note that the plots actually only graph the absolute errors, because the distances $h$ of the points to the identity line are proportional to the the dis-

***Figure 4.29*** – *BOT histograms for calibration assessment of the leave-one-out prediction in the three random effects models for the BMI subsample data. The predictive distributions were estimated with the exact (upper row) and the approximate (lower row) sampling schemes.*



**(a)** Model 1, exact sampling **(b)** Model 2, exact sampling **(c)** Model 3, exact sampling

**(d)** Model 1, approximate sampling **(e)** Model 2, approximate sampling **(f)** Model 3, approximate sampling

tances of the x- and y-coordinates: $h = |x - y| / \sqrt{2}$. The mean relative errors are 0.019, 0.026 and 0.03. So also for the energy scores, the approximation works better in model 1 than in model 2 and model 3.

In Table 4.13 the mean scores are compared, and contrasted with the DIC values. The energy scoring rule ranks model 2 and model 3 equal and model 1 worst. If we approximate the scores, we get lower mean energy scores, with model 2 and model 3 being ranked almost equal. Model 1 is still the worst of the three models. The logarithmic scoring rule prefers model 2 over model 3, but the difference is small between these two models. Model 1 is clearly worse. These conclusions are replicated in the approximate mean scores. It is interesting that the DIC gives model 1 the lowest value, which corresponds to the best model. The ranking of the proper scoring rules is actually reversed, because model 2 is ranked worst by the DIC.

***Table 4.13*** – *Mean energy and logarithmic scores for the cross-validated leave-one-out prediction of the BMI subsample data for the three models, under the exact and approximate sampling schemes, as well as the DIC based on the saturated deviance samples reported by `BayesX`.*

| Model criterion | Scheme | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| ES | exact | 3.27 | 2.67 | 2.67 |
| | approximate | 3.20 | 2.60 | 2.58 |
| log-score | exact | 15.71 | 13.35 | 13.39 |
| | approximate | 15.59 | 13.13 | 13.19 |
| DIC | | 916.67 | 1002.77 | 965.21 |

### Results

The BMI subsample data was fitted best by the parametric non-linear model 2. As this model also had an acceptable calibration in the BOT histogram and was ranked best in the exact leave-one-out predictive assessment by the energy and logarithmic scoring rules, we should choose model 2 based on these subsample results. The DIC again reversed the model ranking from the scoring rules.

The approximate cross-validation scheme worked well in this example, too: The bias of the absolute scores was small and the exact scheme's model rankings were preserved. Also the approximate BOT histograms were close to the exact BOT histograms.

## 4.6.2 Data analysis

Now we include all $n = 2043$ children in our analysis, and also try to use all covariate information. After fitting six models to the BMI data (p. 145), we first assess their goodness-of-fit and examine outlying individuals (p. 147), before doing an approximate cross-validation (p. 148).

### Model fitting

The first three models are retained from section 4.6.1. For the second half of the six candidate models we try to improve model 2 and model 3 and include the additional covariates `wgain2y`, `mDiffBMI`, `breast` and `area`: In model 4, the design variables `mDiffBMI` and `breastbreastFeed` are added as ordinary fixed effects to the model 2 configuration, modifying the baseline level of the child's BMI. The variables `wgain2y` and `areaurban` are added as interaction terms with `age`. For example, a statistically significantly positive coefficient for the latter interaction would then be interpreted as a larger increase of BMI over the first six years of life for children living in an urban area, compared with children living in rural areas. In model 5, the same fixed effect terms are added, this time to the model 3 configuration. In order to smooth the employed P-spline more strongly, we now specify lower hyperparameters $a = b = 0.0001$ for the corresponding variance parameter, as we suggested earlier. In model 6, we extend model 4 by using all three and four dummy variables from the covariates `tvpc` and `mEdu`, respectively, instead of just the two binary variables `tvpcMoreThan1` and `mEduHigh`.

For each of the models, we let `BayesX` Gibbs sample Markov chains of length 200 000. In order to keep the memory load at a manageable size, we saved only every 40-th sample after a burn-in of 100 000 iterations. The result is a sample size of 2500, which should still be large enough to keep the Monte Carlo errors low. This required computing times of 2160, 2117, 2148, 2563, 2568 and 2770 seconds for the six models, respectively. As already observed in the Markov chains for the case study in section 4.6.1, we see high auto-correlations between coefficients samples for the fixed effect of `age`. In the last three models we also note strong negative correlations between `age` and `wgain2yAge`, and to a lower extent between `age` and `areaurbanAge`. This is comprehensible, as the linear predictor part with `age` is

$$\big(\alpha_{\texttt{age},i} + \beta_{\texttt{age}} + \beta_{\texttt{wgain2yAge}}\texttt{wgain2y}_i + \beta_{\texttt{areaurbanAge}}\,\mathbb{I}(\texttt{area}_i == \texttt{urban})\big) \cdot \texttt{age}_{ij}$$

for observation $j$ from child $i$. Since the weight gain until 2 years (covariate `wgain2y`) is always positive, a larger fixed effect $\beta_{\texttt{age}}$ can be balanced to a certain degree by a smaller interaction effect $\beta_{\texttt{wgain2yAge}}$ to retain a similar level of the coefficients sum. If one worried about these posterior correlations, one could try centering the covariate `wgain2y` to "decorrelate" the coefficients. Here we are not interested in the single coefficients samples, but only in the whole age trend, and so do not have problems with the correlations.

The estimated fixed effects of `age` according to the six different models are depicted in Figure 4.31. For the last three models, we set the continuous variables `wgain2y` and `mDiffBMI` to the data point means 8.91 and 5.12, and also the (originally binary) design variables `breastbreastFeed` and `areaurban` to the means 0.6 and 0.79, respectively. This shall ensure that the plots are comparable with the plots from the first three models, which do not include the four covariates `wgain2y`, `mDiffBMI`, `breast` and `area`. The forms of the trends from model 1 in panel (a) and from model 2 in panel (b) are similar to the subsample results in Figure 4.27. Due to the increased number of observations, the credible intervals are much narrower here. This is also the case for model 3 in panel (c), with the mean curve now being smoother than for the subsample.The bump between age 2 and 4 can probably be explained by the local fitting of the B-spline bases, because there are almost no observations around the age of 3 years in the data set. It is instructive that the model 5 fit in panel (e) is indiscernible from the model 3 fit: This means that the different

P-spline variance prior is outweighed by the large number of data points, so that we do not see a clear difference between both trends. Probably only a decreased flexibility of the spline with a lower number of knot locations would have a visible effect on the `age` trend. Panel (d) with the model 4 fit is indiscernible from panel (b). This is actually the check that the adjustment with the means of the four additional covariates works. The mean trend is also very similar to the model 6 trend in panel (f). Because in model 6, the effects of the time-constant covariates `mEdu` and `tvpc` are modelled as 7 dummy variables instead of only 2, the posterior uncertainty about the baseline level for the time trend is larger which is reflected by the wider credible intervals.

### Goodness-of-fit assessment

We plot the posterior-predictive BOT histograms in Figure 4.32. Model 1 provides the worst fit to the data, as the histogram in panel (a) is less left-skewed than the other histograms. Both P-spline models are able to fit the data more closely, if we judge the goodness-of-fit by their histograms in panels (c) and (e). Models 2, 4 and 6 with the nonlinear parametric time trends have the most left-skewed BOT histograms in panels (b), (d) and (f). It is not clear which one of the three fits best.

Analogously to defining outlying individuals as individuals with a high posterior-predictive energy score (cf. page 123), we can look at the posterior-predictive mean energy scores of the models to get numbers for the overall model fits. We also include the mean logarithmic scores in Table 4.14. Both scoring rules assign model 2 the best fit, followed by the other two parametric models 6 and 4, the P-spline models 3 and 5 and finally the simple model 1. The table also shows the posterior expected saturated deviance for the models, which is a traditional goodness-of-fit criterion, see e. g. Spiegelhalter, Best, Carlin, and van der Linde (2002, p. 601). Model 2 has the lowest mean deviance, so that it fits best also according to this measure. It is followed by models 1, 6, 3, 5 and 4. It is interesting that the simple model 1 is ranked second by the deviance, but last by the scoring rules.

***Table 4.14*** – *Posterior-predictive mean energy and logarithmic scores of the six models, as well as the posterior expected saturated deviance.*

| Fit criterion | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| ES | 2.59 | 1.54 | 1.84 | 1.56 | 1.87 | 1.55 |
| Log-score | 14.50 | 10.42 | 11.77 | 10.46 | 11.84 | 10.45 |
| Deviance | 17312.25 | 17310.30 | 17316.27 | 17319.09 | 17319.07 | 17314.14 |

We show the 8 observations having a posterior-predictive BOT value of less than 0.01 in all of the six models in Figure 4.33. The children with IDs 91083394, 94089313 and

94182011 feature untypical BMI jumps at the ages of one and two years. These could be described as truly outlying trajectories. It would be interesting to remove the three children from the statistical analysis, to check if the results are strongly influenced by them. The remaining five children seem to have more normal BMI time series, which are too complicated to be acknowledged by the six models. IDs 91080697, 92082611, 92185191 and 95086051 have a very sharp BMI rise in the first life months, which is not followed adequately by the models. For ID 91982761, the BMI rise occurs too late with regard to the model, so that the third measurement is far below the fitted means, and leads to small posterior-predictive density ordinates and BOT values.

### Cross-validation assessment

The approximate leave-one-out BOT histograms for the six BMI models are shown in Figure 4.34. Since the histograms summarize $n = 2043$ observations here instead of $n = 100$ in Figure 4.29, the appearances are more regular. However, the "calibration message" of the histograms is similar: All models are quite well calibrated, with model 3 in panel (c) and model 5 in panel (e) perhaps being calibrated slightly worse than the other models.

Again in this data example, we observe the pattern of approximate leave-one-out BOT histograms with the first bar for the bin $[0, 0.1]$ being larger, and the rest being left-skewed with the last bar for the bin $(0.9, 1]$ being largest. The pattern is much less pronounced here than in Figure 4.21 for the CD4 data, but it is recognizable.

If we are concerned about the model calibration at the individual observations level, we can inspect the approximate scalar-PIT histograms for the six BMI models. They are shown in Figure 4.35. For all models, there are too few small scalar-PIT values below 0.1 compared to uniform histograms. This means that the lower tails of the scalar predictive distributions are rather too heavy, because too few observations materialize in the lowest parts of the distributions. The model 1 histogram in panel (a) looks worst. The model 3 and model 5 histograms in panels (c) and (e) are better, but still have too large bars for the bin $(0.95, 1]$. This is removed in the remaining panels for the nonlinear parametric time trends models, who still suffer from the too heavily left tailed predictions.

We are interested if the approximate leave-one-out proper scoring rules yield the same result as the DIC. In Table 4.15 the mean energy and logarithmic scores of the models in question as well as the DIC values are shown. Starting with the logarithmic scoring rule, model 6 is ranked as the best model, closely followed by model 4, then models 2, 5, 3 and finally model 1. So the log-scores prefer the nonlinear parametric models, from which the most complex is ranked highest. The energy scoring rule also ranks model 6 best, together with the P-spline model 5. Model 4 is only slightly worse, models 2 and 3 are ranked equally and model 1 is again at the last place. Thus the proper scoring rules agree that the simple model 1 with the linear TP-spline time trend is the worst of all six models. By contrast, the DIC is lowest for model 1, so the DIC ranks model 1 as the best of all six models. The P-spline models 5 and 3 get the second and third places, respectively. The nonlinear parametric models 6, 4 and 2 share the last places. This result is analogous to the result for the subsample data, where Table 4.9 showed that the DIC preferred the simple over the P-spline and the nonlinear parametric model, while the ranking was reverse for the proper scoring rules.

***Table 4.15*** – *Approximate mean energy and logarithmic scores for the cross-validated prediction of the six models, as well as the DIC based on the saturated deviance samples reported by `BayesX`.*

| Model criterion | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---:|---:|---:|---:|---:|---:|
| ES | 3.11 | 2.62 | 2.62 | 2.53 | 2.52 | 2.52 |
| Log-score | 15.38 | 13.01 | 13.35 | 12.87 | 13.21 | 12.85 |
| DIC | 18963.6 | 21392.1 | 19963.2 | 21164.4 | 19643.6 | 21160.4 |

#### Results

From the total six models fitted to the whole BMI data set, models 2, 4 and 6 provided the best fit to the data. This was stated both by the posterior-predictive BOT histograms and the posterior-predictive scoring rules. By contrast, the posterior expected deviance also favoured the simple parametric model 1, besides model 2.

The leave-one-out calibration was better for model 2, 4 and 6 than for the other three models, if we judge this by means of the approximate leave-one-out BOT histograms. The approximate mean scores, which also consider the sharpness of the leave-one-out predictions, rank model 6 best, followed by model 4 and model 2 in the logarithmic scoring rule and model 5 and model 4 in the energy scoring rule. Thus, we would choose model 6 from all models, ignoring that the DIC ranks the oversimplistic model 1 best.

### 4.6.3 Final model

It is interesting if the P-spline can be combined with the nonlinear parametric function (4.6.1) into a single model. We examine a model which is based on model 5 from section 4.6.2 with a P-spline for the fixed effect of `age`. It features not only a random slope, but also the nonlinear part from function (4.6.1) to include the whole parametric function $f(\texttt{age})$ as random effect. This form of individual departure from the population trend could be useful to better fit some outliers from Figure 4.33. Furthermore we take only 7 instead of 8 knots for the P-spline, to smooth the population trend slightly stronger.

We produced a total of 2500 parameter samples for this model specification by thinning out a Markov chain of length 200 000 and discarding a burn-in of 100 000 iterations, within 2898 seconds. The posterior summaries of the fixed effects are tabulated in Table 4.16. We see significant positive associations of male sex, mother's BMI at pregnancy begin, mother's BMI gain during pregnancy and breast feeding with the BMI level. Also the more hours spent outdoors at the age of 4 years, the higher is the BMI level of the child, according to the model. The posterior mean estimate for `wgain2y` can be interpreted as if the child gained one kilogram more weight until the age of 2 years, than the BMI would rise additional 0.09 points per year. By contrast, the association of an urban study center with the BMI is slightly negative. The 95% HPD interval ends near zero for `areaurbanAge` and for `mEduHigh`, so these correlations are only borderline significant. It is even more uncertain if maternal smoking or TV/computer usage is associated with the child's BMI trajectory.

| Coefficient | Mean | Median | SD | lower | upper | Positive |
|---|---|---|---|---|---|---|
| `sexmale` | 0.11 | 0.11 | 0.04 | 0.02 | 0.19 | 0.99 |
| `mBMI` | 0.05 | 0.05 | 0.01 | 0.04 | 0.06 | 1.00 |
| `mDiffBMI` | 0.11 | 0.11 | 0.01 | 0.09 | 0.14 | 1.00 |
| `mSmokeyes` | −0.07 | −0.07 | 0.06 | −0.19 | 0.06 | 0.15 |
| `mEduHigh` | −0.08 | −0.08 | 0.05 | −0.18 | 0.00 | 0.03 |
| `tvpcMoreThan1` | 0.04 | 0.04 | 0.05 | −0.06 | 0.14 | 0.77 |
| `outdoor` | 0.05 | 0.05 | 0.02 | 0.01 | 0.08 | 1.00 |
| `breastbreastFeed` | 0.17 | 0.17 | 0.05 | 0.08 | 0.26 | 1.00 |
| `wgain2yAge` | 0.09 | 0.09 | 0.00 | 0.08 | 0.10 | 1.00 |
| `areaurbanAge` | −0.03 | −0.03 | 0.01 | −0.05 | 0.00 | 0.03 |

***Table 4.16*** – *Posterior summaries for fixed effects coefficients in the final model: In addition to the posterior mean, median and standard deviation of the coefficient, the lower and upper bound of the 95% HPD-interval and the posterior probability that the coefficient is positive are shown.*

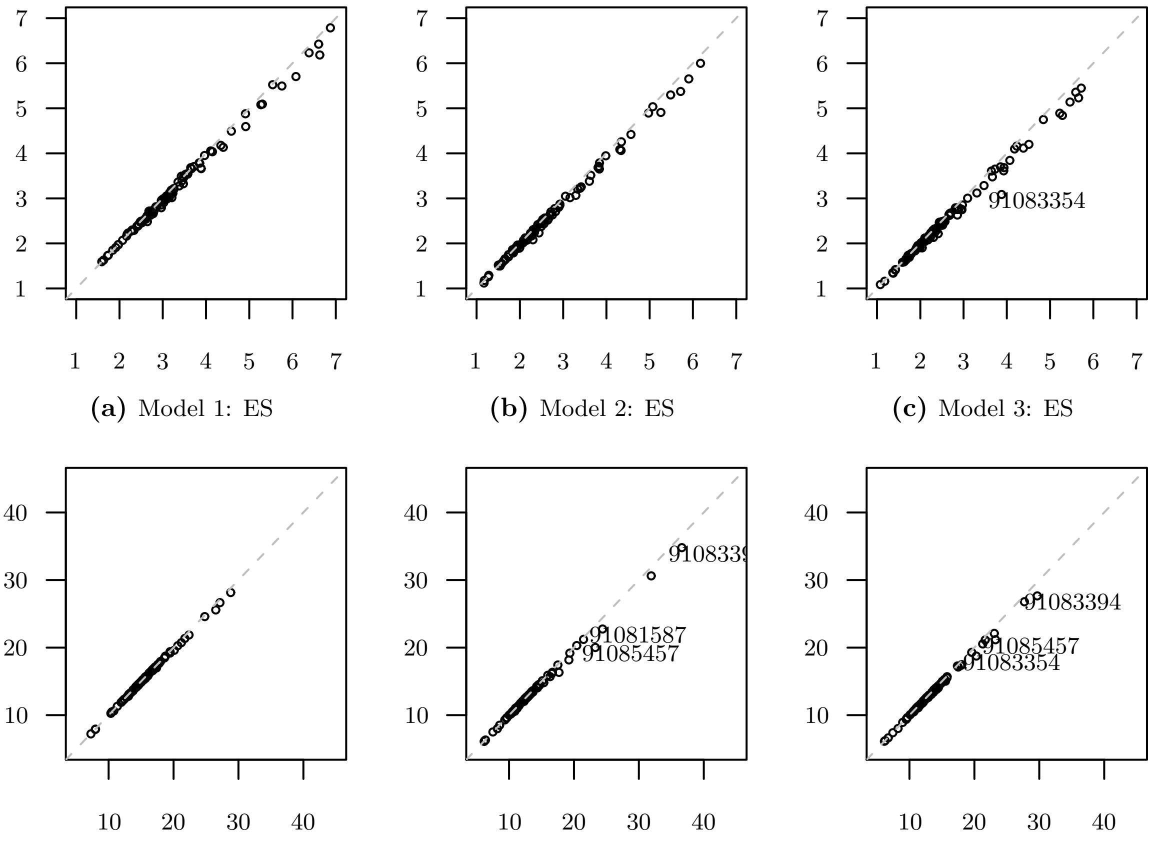
The fixed effect age trend estimate is graphed in Figure 4.36. The trend is noticeably smoother than in panel (e) on page 153. This is supposedly due to the use of 7 instead of 8 knot locations. Yet, the overall picture has not changed much, only after 6 years (where few data points are observed and the uncertainty is large) the mean curve differs from the original model 5 curve.

The mean deviance is 17321.7, the mean posterior-predictive log-score and energy score are 10.6 and 1.59, respectively. While this is the worst mean deviance of all models, the scores are almost as good as for the models 2, 4 and 6 with fixed nonparametric age trend (cf. Table 4.14 on page 147). The posterior-predictive BOT and PIT histograms are shown in Figure 4.37. The BOT histogram in panel (a) attests the new model a better fit than the old model 5, with panel (e) on page 154. The PIT histogram in panel (b) reinforces this conclusion.

In order to check the leave-one-out calibration of the final model, we plot the approximate cross-validation BOT and PIT histograms in Figure 4.38, as the exact cross-validation would have required ca. 1645 hours. The predictive calibration looks very good in the BOT histogram in panel (a) compared with the histograms in Figure 4.34 on page 156. The PIT histogram in panel (b) is good too, but has similar defects to the other models' PIT histograms in Figure 4.35 on page 157.

The DIC is 21103.7, and the approximate cross-validation log-score and energy score are 12.95 and 2.54, respectively. That ranks the new model between the old P-spline models and the nonlinear parametric models with respect to the DIC and the log-score. The energy score is the fourth best of all seven models which have been examined.

***Figure 4.30*** – *Comparison of exact and approximate scores for leave-one-out prediction in the three random effects models (columns) for the BMI subsample data. The panels in the upper row compare the energy scores (ES), while the panels in the lower row compare the log-scores. Individuals where the absolute difference between the exact and approximate score values exceeds 0.5 (ES) or 1.5 (log-scores) are labelled.*



**(a)** Model 1: ES

**(b)** Model 2: ES

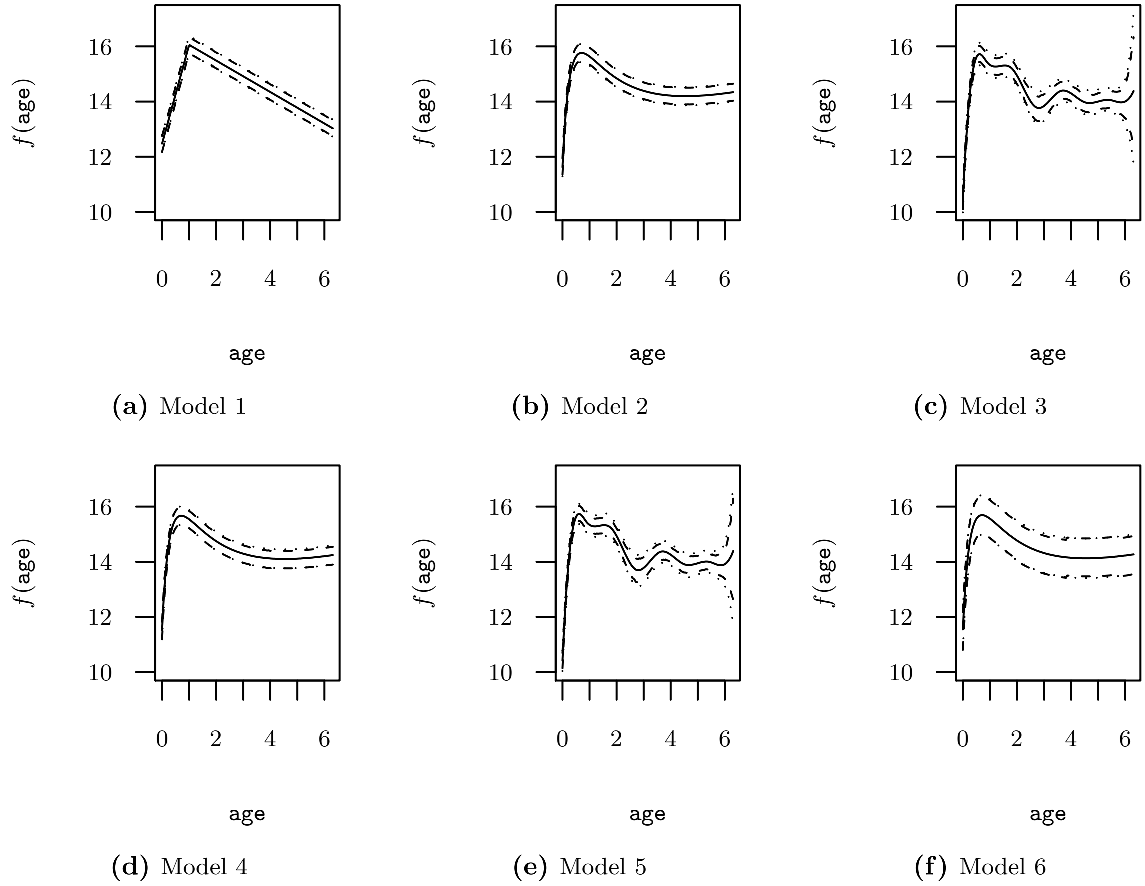
**(c)** Model 3: ES

**(d)** Model 1: log-score

**(e)** Model 2: log-score

**(f)** Model 3: log-score

***Figure 4.31*** – *Estimated fixed effects time trends (including the intercept) in the six models: Means (——), pointwise HPD (- - -), and simultaneous (........) credible intervals at the 0.95 level are plotted. For models 4–6, the time trends samples which were averaged include the sampled effects of the covariates `wgain2y`, `mDiffBMI`, `breast` and `area` at their data point means.*



**(a)** Model 1

**(b)** Model 2

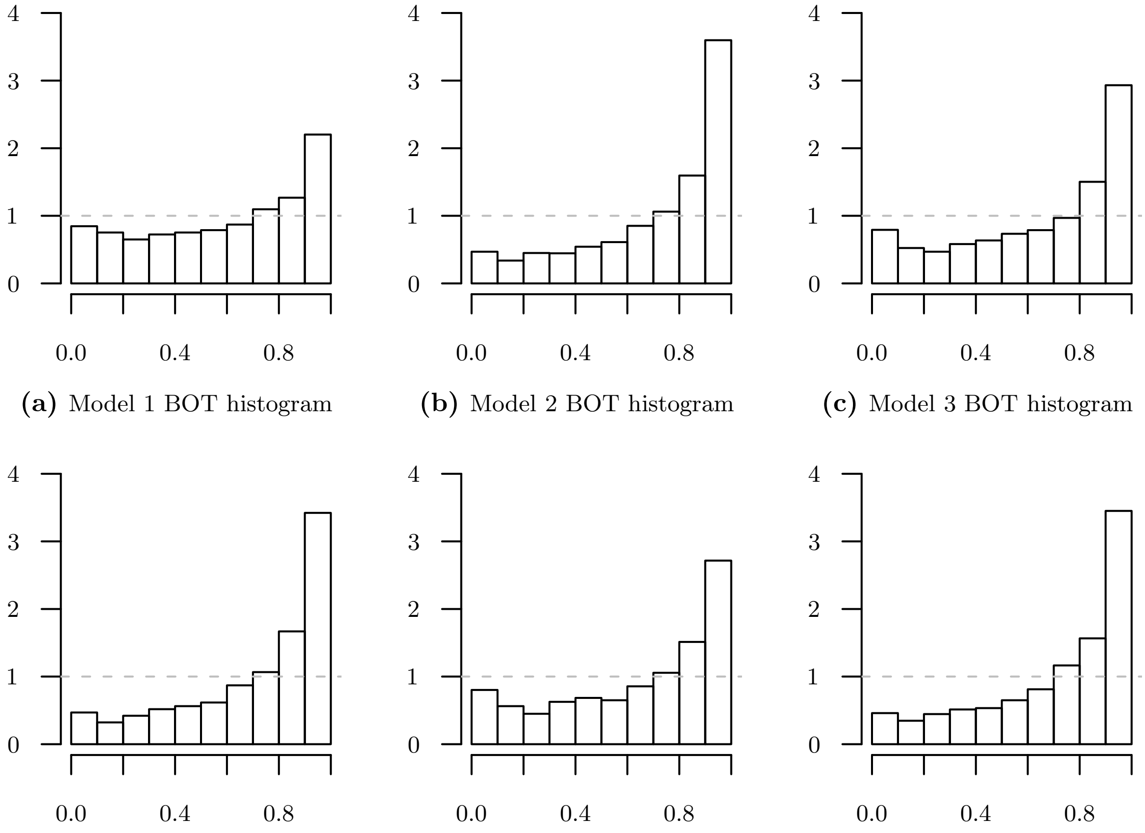
**(c)** Model 3

**(d)** Model 4

**(e)** Model 5

**(f)** Model 6

***Figure 4.32*** – *Posterior-predictive BOT histograms for goodness-of-fit assessment of the six models.*



**(a)** Model 1 BOT histogram

**(b)** Model 2 BOT histogram

**(c)** Model 3 BOT histogram

**(d)** Model 4 BOT histogram

**(e)** Model 5 BOT histogram

**(f)** Model 6 BOT histogram

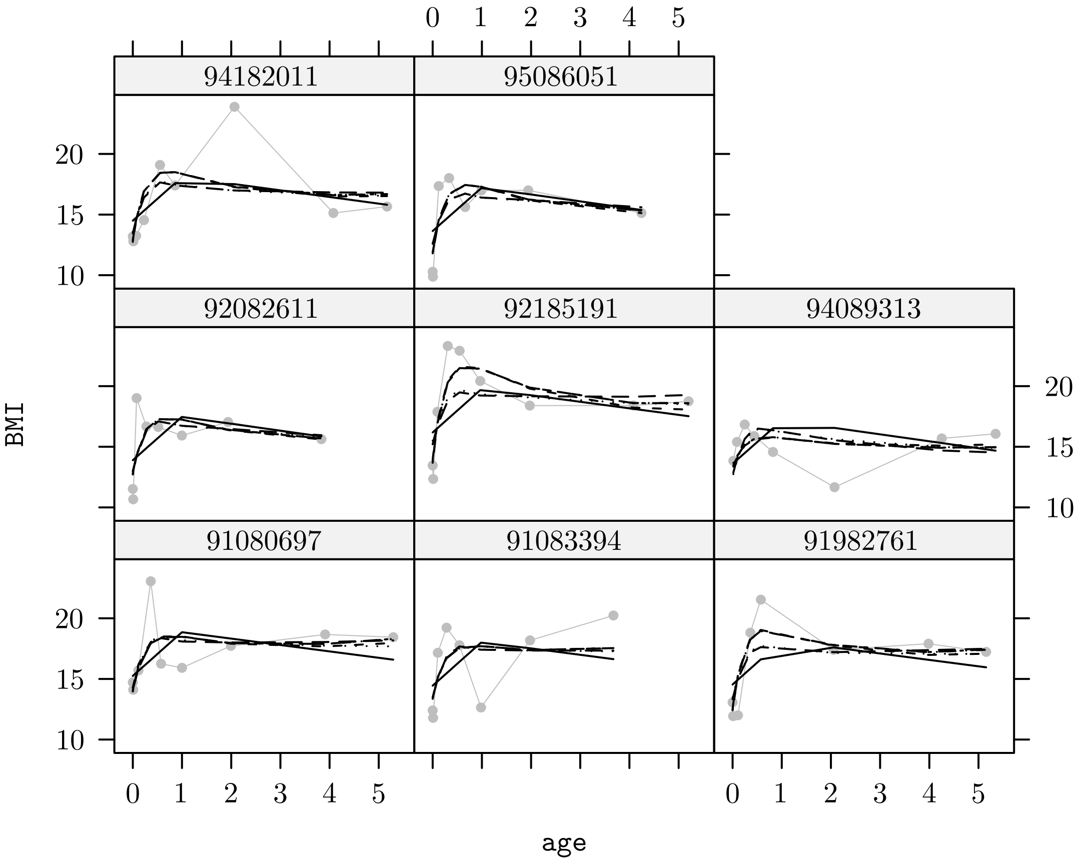
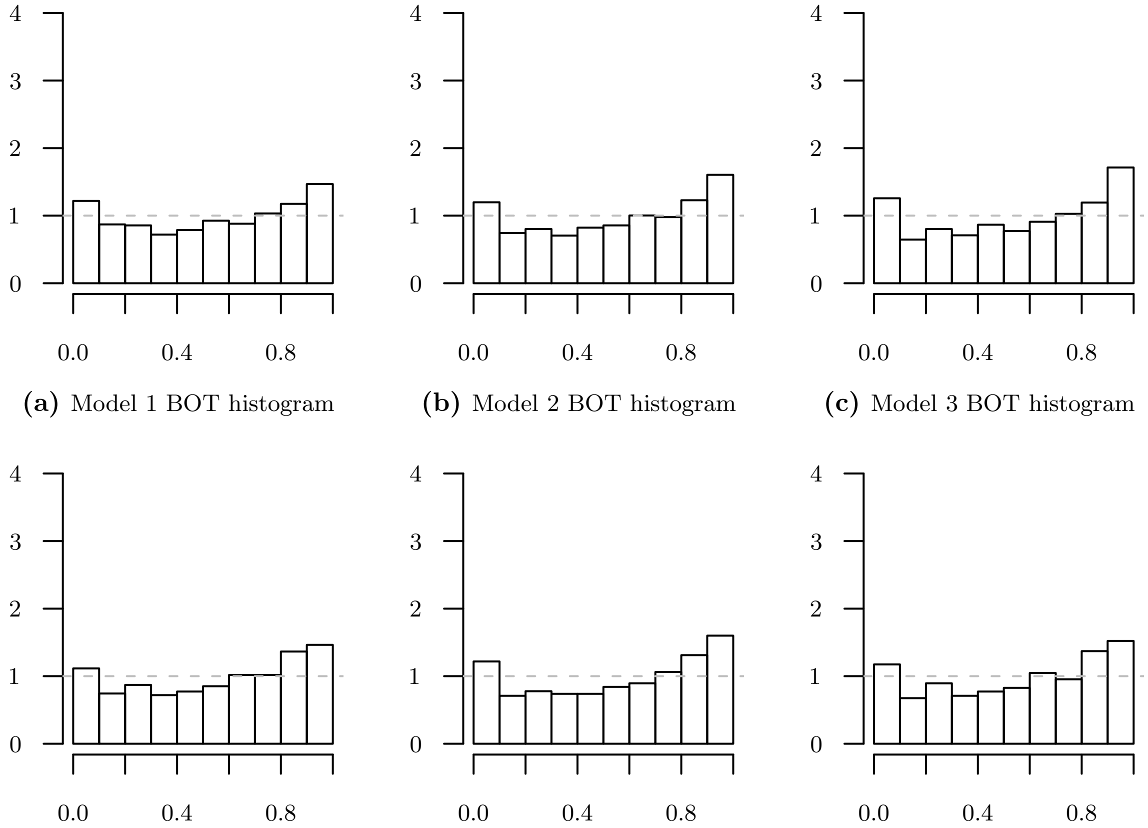


***Figure 4.33*** – *Model fits (estimated posterior means) for the children with posterior-predictive BOT values less than 0.01 in all of the six models, which are discerned by line type: model 1 (____), model 2 (_ _ _ ), model 3 (........), model 4 (._._), model 5 (____), and model 6 (.__._).*

***Figure 4.34*** – *Approximate BOT histograms for calibration assessment of the leave-one-out predictive distributions implied by the six models.*



**(a)** Model 1 BOT histogram

**(b)** Model 2 BOT histogram

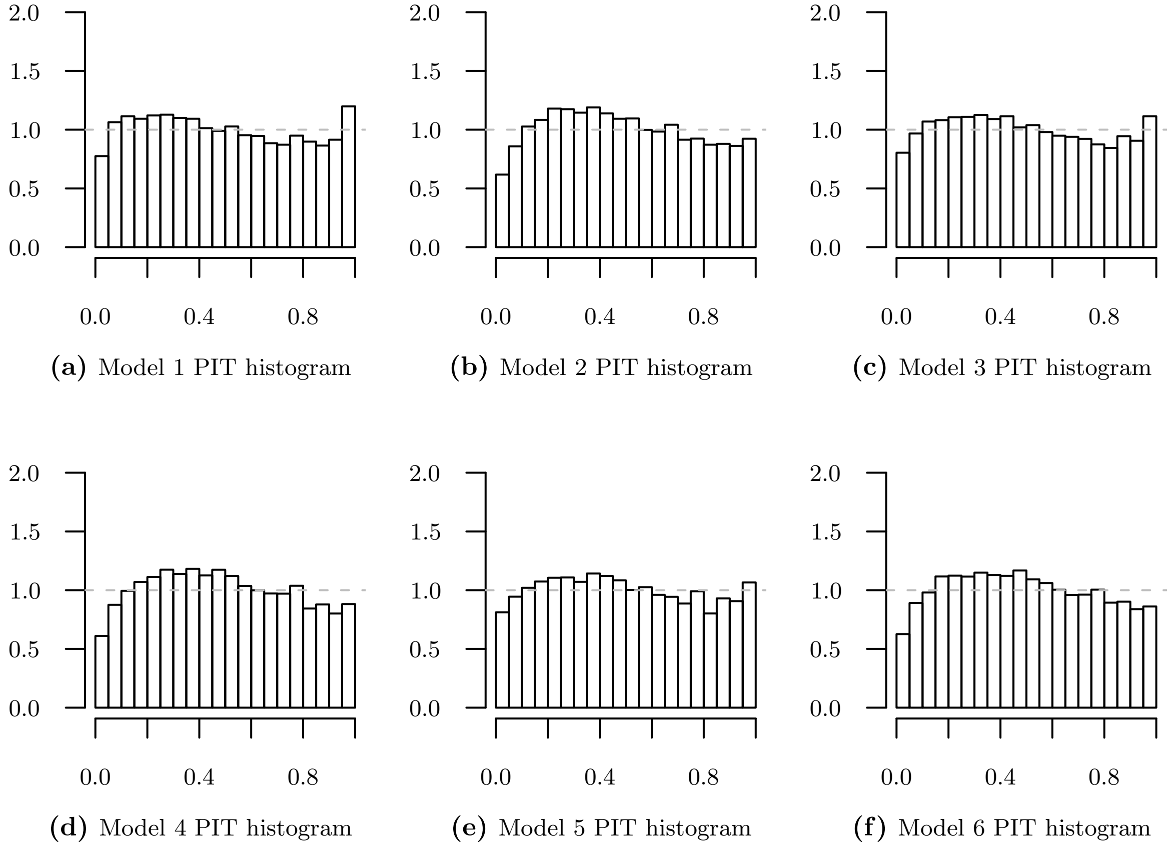
**(c)** Model 3 BOT histogram

**(d)** Model 4 BOT histogram

**(e)** Model 5 BOT histogram

**(f)** Model 6 BOT histogram

***Figure 4.35*** – *Approximate scalar-PIT histograms for cross-validated calibration assessment of the scalar predictive distributions in the six models.*



**(a)** Model 1 PIT histogram

**(b)** Model 2 PIT histogram

**(c)** Model 3 PIT histogram

**(d)** Model 4 PIT histogram

**(e)** Model 5 PIT histogram

**(f)** Model 6 PIT histogram

***Figure 4.36*** – *Estimated fixed effects time trend in the final model: Means (——), pointwise HPD (- - -), and simultaneous (........) credible intervals at the 0.95 level are plotted. The trends samples which were averaged include the sampled effects of the covariates* `wgain2y`, `mDiffBMI`, `breast` *and* `area` *at their data point means.*

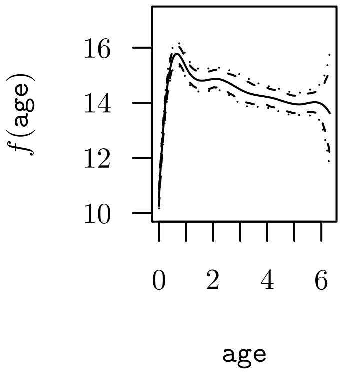
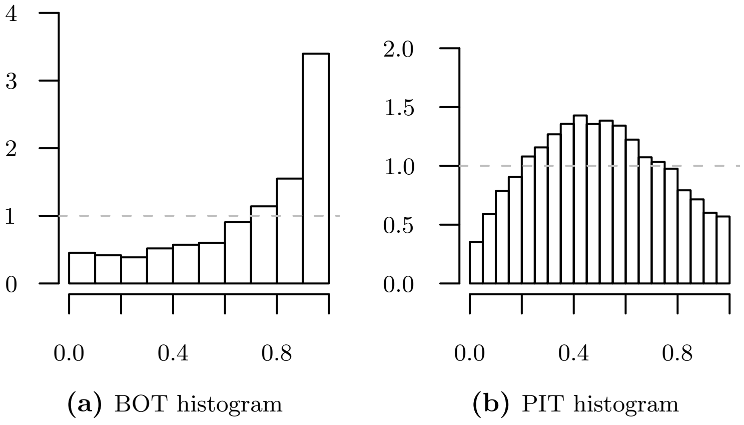


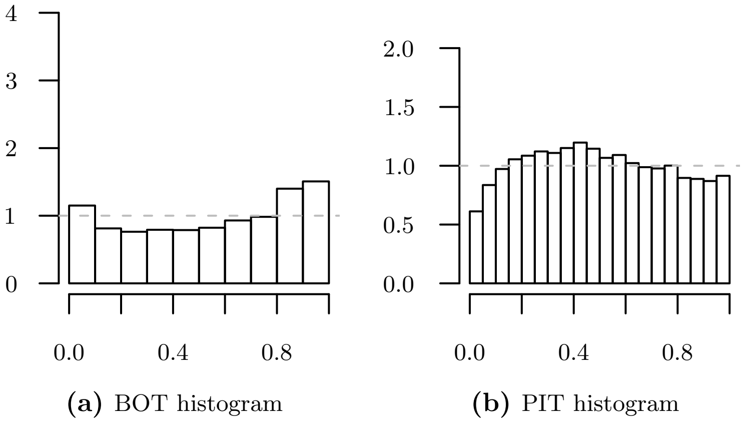
***Figure 4.37*** – *Posterior-predictive BOT and PIT histograms for goodness-of-fit assessment of the final model.*



**(a)** BOT histogram

**(b)** PIT histogram

***Figure 4.38*** – *Approximate BOT and PIT histograms for leave-one-out predictive calibration assessment of the final model.*



**(a)** BOT histogram

**(b)** PIT histogram

## 4.7 Summary

In this chapter, we have applied the Marshall-Spiegelhalter approach to Bayesian random effects models, which have been estimated with the MCMC implementation of `BayesX`. The cross-validation assessment for the longitudinal data was understood as a leave-one-out loop on the level of the individuals, as opposed to the level of single observations. For a simulated data set and two real data examples, we have compared the exact cross-validation results with the Marshall-Spiegelhalter approximations, and found that the results were very close to each other: the ranking of the models by means of the scoring rules was mostly preserved by the approximate scheme, and the bias of the absolute scores was small, especially for the logarithmic scoring rule. Also the calibration results were essentially retained under the approximate scheme, although the approximation of single BOT values and resulting BOT histograms was more difficult than the score approximation.

Moreover, we have experienced that the goodness-of-fit assessment of the longitudinal models can be based on the posterior-predictive model scores. This usage of posterior-predictive samples should be favoured over the production of BOT or PIT histograms, due to the easier and objective interpretability and ranking of the models' results. A related topic is the detection of outlying individuals, which we understood as individuals which cannot be fitted well by the model. For this task, the BOT values can be utilized, with smaller BOT values indicating a worse fit. While the energy score values can also be used, the single posterior-predictive logarithmic scores should be interpreted carefully, especially when the dimensions $n_i$ of the time series vary.

In comparison with the current default model criteria printed by `BayesX`, namely the posterior expected deviance and the DIC, the approximate model scores were competitive: for the simulated data, the DIC could not identify the correct model as the best model, in contrast to the approximate scores. Also for the BMI data set, the DIC preferred a too simple model, judging both from the goodness-of-fit of this model and its approximate leave-one-out scores. The posterior expected deviance, too, seems to have a tendency to prefer (too) simple models, as the BMI data analysis suggests. These findings motivate an integration of the approximate leave-one-out cross-validation scheme into `BayesX`, so that the users would also be provided with the posterior-predictive and approximate leave-one-out energy and logarithmic scores of the fitted model as an interesting supplement to the current output.

# 5 Summary and Discussion

In this thesis, we have implemented exact and approximate predictive assessment for two Bayesian hierarchical model classes: While the conjugate change point models are applicable to single time series data, the random effects models are used to analyze longitudinal data. For the time series, two natural assessment schemes were used, which assess either the quality of one-step-ahead or leave-one-out predictions. For the longitudinal data, only the leave-one-out scheme is sensible on the level of individuals. However, for data sets with equally long time series, one could also imagine a one-step-ahead scheme which works on the level of individual observations. The predictions and the materialized observations were compared with PIT/BOT transformations and proper scoring rules to address both the calibration and the sharpness of the predictions. These evaluations of the predictive distributions were fully based on samples (either directly from the forecasters or from their underlying distributions), and can thus be applied to a wide range of situations.

For both model classes, the implemented approximations of the Marshall-Spiegelhalter type worked well: Although always an optimistic bias of the approximate scores and PIT values could be observed, the conveyed statement of the exact results was often retained by the approximate results. The advantage of the approximations is entirely computational, but this advantage can be vital for large data sets. The approximate predictive assessment schemes should thus be utilized when the exact assessment is not feasible any longer, as was the case for the genetic data in section 3.7 or the full CD4 and BMI data sets in sections 4.5 and 4.6, respectively. For the latter two data examples, we selected a small subset of the data to be able to compare with the approximate with the exact results, at least for a smaller part of the data. This strategy could be generalized in statistical practice: test the approximate scheme on a small subsample of the data, and if it yields satisfactory approximations to the exact scheme, apply the approximate scheme to the whole data set, where the exact scheme is infeasible.

An important point of this thesis is the contrasting of the exact and approximate predictive assessment results with the analogous posterior-predictive results. We have shown that the easily accessible posterior-predictive results are suitable for, and only for, the goodness-of-fit assessment of the considered models. This comprises the detection of poorly fitted observations, by comparing the posterior-predictive distributions with the corresponding known observations. However, in general the posterior-predictive results are far away from the corresponding exact and approximate predictive assessment results: the sampling scheme is just much too optimistic about the forecasting capabilities of the models. The scores are thus shrunk to 0, the PIT values are shrunk to 0.5 and the BOT values are shrunk to 1. Moreover, it is crucial to remember that a model which fits the known data well is not guaranteed to be a good forecaster for yet unknown data. This was shown very clearly in the simulated data analysis with the random effects model in section 4.4. That is the fundamental problem of overfitting models, and many examples were given in the case studies in sections 3.4–3.6.

To conclude, we propose two possible extensions of topics covered in this thesis.

First, note that due to the definition of our change point model in section 3.2.2, there is a shortcoming of the modelling approach. It is relevant for the Tokyo rainfall data case study in section 3.5.2 and was neither avoided by Kitagawa (1987), who introduced the data set into the literature: There is no implemented connection between the rain probabilities on 31st December and 1st January. However, since the data is a “cyclic” time series, it would be better to somehow penalize probability trend differences between these two adjacent days. To remedy this shortcoming, our change point model could be extended to cyclic time series. Essentially the blocks $\boldsymbol{y}_{[1,\theta_1]}$ before the first change point and $\boldsymbol{y}_{(\theta_k,n]}$ after the last change point would have to share the *same* model parameter $\boldsymbol{\xi}^{(1)} \equiv \boldsymbol{\xi}^{(k+1)}$. Yet, note that *two* change points are necessary in this cyclic time series model to distinguish two seasons in the year: if $k = 1$, then the “first” and the “last” block would still share the same parameter. In order to allow a change point between the last observation $y_n$ and the first observation $y_1$, an *optional* change point at time $n$ could be introduced. From the other perspective, the normal time series structure is a special case of the cyclic time series, where there is a fixed change point between the end and the start of the time series. Moreover, the cyclic time series framework could be extended to seasonal time series, which cannot be summarized into a single cyclic time series as it was the case for the Tokyo rainfall data. For example, instead of the rainy days we could have recorded temperature measurements. Then the model parameters for the seasonal trend could still form a cyclic time series, and each observation would be assigned the appropriate parameter via the calendar day.

Second, the sampling based evaluation of the predictive distributions offers a very easy possibility to consider transformations of the independent variable. We will illustrate that point with the CD4 data from section 4.5, where the direct modelling of the untransformed counts with a normal linear mixed model posed problems – the calibration of the considered models was unsatisfactory. A conventional transformation of the CD4 counts is the square root transformation, which is the variance-stabilizing transformation for Poisson-distributed count data. So we could transform the counts, and consider models for the root-counts. These would produce predictive samples for the square root of the CD4 counts. In order to compare these new models with the old models on the original scale, we could then just square the root-count predictive samples to map them onto the original count scale. The squared samples would then produce BOT, PIT, and proper scoring rule values which could be compared directly with the old models' results. Analogously, we could try to fit the natural logarithm of the CD4 counts, to avoid the problem that the normal likelihood includes impossible negative counts in the inference. These models could be compared directly with the other models on the original counts scale, too, by exponentiating the predictive log-counts samples of the log-CD4 models. To mention an example for the conjugate change point models, in section 3.7 we could try to model the logit transformed GC proportions instead of the original proportions, which could lead to a better normal approximation of the observations conditional on the means.

# A Appendix

## A.1 CRPS formula

We will show that for $x \in \mathbb{R}$ and $Y, Y^* \overset{iid}{\sim} G$ with finite expectation $\mathbb{E}(Y) = \int y \, dG(y)$ the identity

$$\mathbb{E}\,|Y - x| - \frac{1}{2}\mathbb{E}\,|Y - Y^*| = \int \left\{G(y) - \mathbb{I}_{[x,+\infty)}(y)\right\}^2 \, dy \tag{A.1.1}$$

holds. The integration in (A.1.1) and hereafter extends over the entire real line.

The proof follows lemmas 2.1 and 2.2 of Baringhaus and Franz (2004, p. 192), which start with two independent random variables $X \sim F$ and $Y \sim G$ with finite expectations. First note the basic identity

$$\begin{aligned} |x - y| &= \begin{cases} x - y = \int \mathbb{I}_{[y,x)}(u) \, du & \text{if } x \geq y \\ y - x = \int \mathbb{I}_{[x,y)}(u) \, du & \text{if } x < y \end{cases} \\ &= \int \mathbb{I}_{[y,x)}(u) + \mathbb{I}_{[x,y)}(u) \, du, \end{aligned}$$

which is true because $[x, y) = \emptyset$ and thus $\mathbb{I}_{[x,y)}(u) \equiv 0$ if $x \geq y$ (and analogously for $x < y$). Therefore the expected distance between $X$ and $Y$ can be rewritten as

$$\begin{aligned} \mathbb{E}\,|X - Y| &= \mathbb{E} \int \mathbb{I}_{[Y,X)}(u) + \mathbb{I}_{[X,Y)}(u) \, du \\ &= \int \mathbb{E}\,\mathbb{I}_{[Y,X)}(u) + \mathbb{E}\,\mathbb{I}_{[X,Y)}(u) \, du \\ &= \int \mathbb{P}(Y \leq u < X) + \mathbb{P}(X \leq u < Y) \, du \\ &= \int \mathbb{P}(Y \leq u) \, \mathbb{P}(u < X) + \mathbb{P}(X \leq u) \, \mathbb{P}(u < Y) \, du \\ &= \int G(u)(1 - F(u)) + F(u)(1 - G(u)) \, du, \end{aligned} \tag{A.1.2}$$

where we have used Fubini's theorem for the change of integration and expectation order, and the stochastic independence of $X$ and $Y$.

Now introduce two independent copies of $X$ and $Y$, namely $X^* \sim F$ and $Y^* \sim G$.

From (A.1.2) it follows that

$$\begin{aligned}
&\mathbb{E}\,|X-Y| - \frac{1}{2}\,\mathbb{E}\,|X-X^*| - \frac{1}{2}\,\mathbb{E}\,|Y-Y^*| \\
&\quad = \int G(u)(1-F(u)) + F(u)(1-G(u)) - \frac{1}{2}\cdot 2F(u)(1-F(u)) - \frac{1}{2}\cdot 2G(u)(1-G(u))\,du \\
&\quad = \int G(u) - F(u)G(u) + F(u) - F(u)G(u) - F(u) + F(u)^2 - G(u) + G(u)^2\,du \\
&\quad = \int F(u)^2 - 2F(u)G(u) + G(u)^2\,du \\
&\quad = \int (F(u)-G(u))^2\,du. \qquad \text{(A.1.3)}
\end{aligned}$$

If we choose the cdf $F(u) := \mathbb{I}_{[x,+\infty)}(u)$, we arrive at the point-mass-in-$x$ distributed $X, X^* \overset{iid}{\sim} \delta_x$. Therefore we can substitute $x$ for $X, X^*$ in identity (A.1.3), giving

$$\mathbb{E}\,|x-Y| - \frac{1}{2}\,\mathbb{E}\,|Y-Y^*| = \int (\mathbb{I}_{[x,+\infty)}(u) - G(u))^2\,du,$$

because $\mathbb{E}\,|X-X^*| = \mathbb{E}\,|x-x| = 0$. This completes the proof of identity (A.1.1).

## A.2 Saturated Deviance and DIC

The saturated deviance in the model framework from section 4.2.3 is defined as

$$D(\boldsymbol{\xi},\boldsymbol{\alpha}) = 2\log\left\{\frac{g(\boldsymbol{y}\,|\,\sigma^2)}{f(\boldsymbol{y}\,|\,\boldsymbol{\xi},\boldsymbol{\alpha})}\right\}.$$

It compares the likelihood $f(\boldsymbol{y}\,|\,\boldsymbol{\xi},\boldsymbol{\alpha}) = \prod_{i=1}^{n} \mathrm{N}_{n_i}(\boldsymbol{y}_i\,|\,\boldsymbol{\mu}_i, \sigma^2\boldsymbol{I}_{n_i})$ of the parameters to the data density where the means $\boldsymbol{\mu}_i$ are replaced by the actual observations $\boldsymbol{y}_i$, $g(\boldsymbol{y}\,|\,\sigma^2) = \prod_{i=1}^{n} \mathrm{N}_{n_i}(\boldsymbol{y}_i\,|\,\boldsymbol{y}_i, \sigma^2\boldsymbol{I}_{n_i})$, by means of the well-known likelihood ratio statistic.

The Deviance Information Criterion (DIC) was proposed by Spiegelhalter, Best, Carlin, and van der Linde (2002) and is based on the (saturated) deviance. Let

$$p_D := \mathbb{E}\big[D(\boldsymbol{\xi},\boldsymbol{\alpha})\,|\,\boldsymbol{y}\big] - D(\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\alpha}})$$

be the difference between the posterior expected deviance and the deviance at the posterior expected parameter values $\bar{\boldsymbol{\xi}} := \mathbb{E}(\boldsymbol{\xi}\,|\,\boldsymbol{y})$, $\bar{\boldsymbol{\alpha}} := \mathbb{E}(\boldsymbol{\alpha}\,|\,\boldsymbol{y})$. $p_D$ is the effective number of parameters in the Bayesian model. The DIC is then defined in analogy to the AIC as

$$DIC := D(\bar{\boldsymbol{\xi}},\bar{\boldsymbol{\alpha}}) + 2p_D.$$

It is oriented as the proper scoring rules, i. e. lower DIC values correspond to better models.

## A.3 Paired permutation test for difference of mean scores

Suppose we have computed two scores $r_i$ and $s_i$ for the prediction of each observation $y_i$, $i = 1, \dots, n$. In our applications, the score $r_i$ is usually obtained from some model, say $M_r$, predicting the observation $y_i$ from (a subset of) the remaining observations $\boldsymbol{y}_{\mathcal{N}\setminus\{i\}}$, while $s_i$ is obtained from the prediction prescribed by $M_s$. The proper scoring rule which compares the forecasters $F_{r,i}$ and $F_{s,i}$ with the materialized observation $y_i$ is of course the same for both models, e. g. the CRPS and $r_i = CRPS(F_{r,i}, y_i)$, $s_i = CRPS(F_{s,i}, y_i)$.

We want to compare the mean scores $\bar{r}$ and $\bar{s}$ with a formal significance test, in order to examine if their difference $\bar{d} = \bar{r} - \bar{s}$ is statistically significant on a certain level (usually 0.01 or 0.05). Then the paired permutation test provides a convenient solution, because unlike e. g. the paired Student t-test, it does not require distribution assumptions or trust in asymptotic behaviour.

The null hypothesis is that the mean scores $\mu_r$ and $\mu_s$ in the population are equal, $\mu = \mu_r = \mu_s$. The alternative hypothesis is the contrary, $\mu_r \neq \mu_s$. We have estimated the population parameters $\mu_r$ and $\mu_s$ by $\bar{r}$ and $\bar{s}$ from a paired sample of size $n$. The idea of the permutation test is that under the null hypothesis, the values of $r_i$ and $s_i$ could be exchanged without changing the expected means in the two score sets. These would still be $\mathbb{E}(\bar{R}) = \mathbb{E}(\bar{S}) = \mu$. Exchanging the values of the $i$-th pair is equivalent to changing the sign of the difference $d_i = r_i - s_i$. So a permuted test statistic is simulated as $\bar{d}^*_{[b]} = \frac{1}{n}\sum_{i=1}^{n} d_i \cdot (-1)^{z_{i,[b]}}$ where $z_{i,[b]}$ is drawn from a Bernoulli distribution with probability 0.5 for all observations $i = 1, \dots, n$. The randomized permutation is done for $b = 1, \dots, B = 10\,000$, say. Then the two-sided p-value

$$p = \mathbb{P}(|D^*| > |d|) = \mathbb{P}(|D| > |d| \mid H_0 \text{ is true})$$

can be approximated by the Monte Carlo estimate

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|d^*_{[b]}| > |d|).$$

If $n$ is small enough, all $2^n$ permutations can be considered. Baker and Tilbury (1993) have devised a fast algorithm for the approximation of the resulting “exact” p-value with fixed accuracy in polynomial time. See Jolliffe (2007, p. 646) for permutation tests applied to the inference of verification measures in meteorology.

## A.4 Normal-normal mixture

Let $x \mid \mu \sim \mathrm{N}(\mu, 1)$ and $\mu \sim \mathrm{N}(0, 1)$. Then we have from Held (2008, p. 148) that the posterior of $\mu$ is again a normal distribution, namely

$$\mu \mid x \sim \mathrm{N}\left(\frac{1}{2}x, \frac{1}{2}\right).$$

Thus we have from Bayes' theorem that the marginal density of $x$ is

$$\begin{aligned}
f(x) &= \frac{f(x \mid \mu) f(\mu)}{f(\mu \mid x)} \\
&= \frac{\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\} \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\mu^2\right\}}{\frac{\sqrt{2}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \cdot 2(\mu - \frac{1}{2}x)^2\right\}} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left\{-\frac{1}{2}\left[x^2 - 2\mu x + \mu^2 + \mu^2 - 2(\mu^2 - \tfrac{1}{2}\mu x + \tfrac{1}{4}x^2)\right]\right\} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left\{-\frac{1}{2 \cdot 2}x^2\right\} \\
&= \mathrm{N}(x \mid 0, 2).
\end{aligned}$$

# B Nomenclature

AIC *A*kaike's *I*nformation *C*riterion

BIC *B*ayesian *I*nformation *C*riterion

BOT *B*ox *O*rdinate *T*ransform

(C)RPS (*C*ontinuous) *r*anked *p*robability *s*core

cdf *c*umulative *d*istribution *f*unction

DIC *D*eviance *I*nformation *C*riterion

DIC *D*eviance *I*nformation *C*riterion

ES *E*nergy *s*core

iid *i*ndependent *i*dentically *d*istributed

MCMC *M*arkov *c*hain *M*onte *C*arlo

pdf *p*robability *d*ensity *f*unction

PIT *P*robability *I*ntegral *T*ransform

$[s,t]$, $(s,t]$ For integers $s \leq t$, we abbreviate $[s,t] := \{s, s+1, \dots, t\}$ If $s > t$, $[s,t] := \emptyset$. Similarly, for integers $s < t$, define $(s,t] := \{s+1, s+2, \dots, t\}$ and for $s \geq t$ we have $(s,t] := \emptyset$.

$\boldsymbol{y}_{\mathcal{S}}$ For $\mathcal{S} \subset \mathcal{N}$, $\boldsymbol{y}_{\mathcal{S}} := \{y_t \,|\, t \in \mathcal{S}\}$. Therefore, if $\mathcal{S} = \emptyset$, $\boldsymbol{y}_{\mathcal{S}} = \emptyset$.

$\mathcal{N}$ The set of all indexes is $\mathcal{N} := \{1, 2, \dots, n\}$.

$n$ The number of observations (individuals) is denoted as $n$.

$(\boldsymbol{B} \,|\, \boldsymbol{C})$ If $\boldsymbol{B} = (b_{ij}) \in \mathbb{R}^{k \times m}$ and $\boldsymbol{C} = (c_{ij}) \in \mathbb{R}^{k \times n}$ are matrices with the same number of rows ($k$), then $\boldsymbol{A} = (\boldsymbol{B} \,|\, \boldsymbol{C})$ denotes the matrix which is concatenated from the columns of $\boldsymbol{B}$ and the columns of $\boldsymbol{C}$. That is, $\boldsymbol{A} = (a_{ij})$

with elements

$$a_{ij} = \begin{cases} b_{ij} & \text{if } 1 \leq j \leq m \\ c_{i,j-m} & \text{if } m+1 \leq j \leq m+n. \end{cases}$$

$\mathbf{1}_k$ The one-vector of dimension $k \in \mathbb{N}$ is $\mathbf{1}_k := (1, 1, \ldots, 1)'$.

$\boldsymbol{I}_k$ The identity matrix of dimension $k \in \mathbb{N}$ is $\boldsymbol{I}_k := \operatorname{diag} \mathbf{1}_k$.

$\operatorname{diag} \boldsymbol{x}$ For $\boldsymbol{x} \in \mathbb{R}^k$, $\operatorname{diag} \boldsymbol{x}$ is the diagonal matrix with the elements of $\boldsymbol{x}$ arranged on the diagonal:

$$\operatorname{diag} \boldsymbol{x} := \begin{pmatrix} x_1 & & & 0 \\ & x_2 & & \\ & & \ddots & \\ 0 & & & x_k \end{pmatrix}$$

$\delta_x$ The Dirac point measure in $x$ is denoted as $\delta_x$.

$\mathbb{I}_{\mathcal{A}}(x)$ The indicator function for the set $\mathcal{A}$ is $\mathbb{I}_{\mathcal{A}}$ and returns 1 if the argument is an element of $\mathcal{A}$ and 0 else.

$\|\boldsymbol{z}\|$ the Euclidean norm $(\sum_{j=1}^{k} z_j^2)^{1/2}$ of $\boldsymbol{z} \in \mathbb{R}^k$

$\Phi$ The cdf of the standard normal distribution is denoted as $\Phi$.

$O(h(m))$ The notation $O(h(m))$ describes an algorithm with complexity $h(m)$ in the variable $m$. More formally, if $g(m) \geq 0$ is the exact computational run-time given $m$ (e. g. the number of samples), then $g(m) = O(h(m))$ means that $|g(m)| \leq M\,|h(m)|$ for some $M > 0$ and sufficiently large $m$. For example, if $h \equiv \mathrm{id}$ then the algorithm has linear complexity $O(m)$.

$\boldsymbol{\theta}_{[b]}$ We denote the $b$-th sample of some parameter $\boldsymbol{\theta}$ using a parenthesized index as $\boldsymbol{\theta}_{[b]}$ for better distinction from the other indexes.

$y^*$ The replication of $y$ is denoted as $y^*$ – this notation is needed for the posterior-predictive distributions (this is only used for replications with the letter $y$).

# References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Baker, R. D. and J. B. Tilbury (1993). Algorithm AS 283: Rapid computation of the permutation paired and grouped t-tests. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 42*(2), 432–441.

Baringhaus, L. and C. Franz (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis 88*(1), 190–206.

Barry, D. and J. A. Hartigan (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association 88*(421), 309–319.

Belitz, C., A. Brezger, T. Kneib, and S. Lang (2009a). *BayesX Methodology Manual.* version 2.00.

Belitz, C., A. Brezger, T. Kneib, and S. Lang (2009b). *BayesX Reference Manual.* version 2.00.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science 10*, 3–66.

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General) 143*(4), 383–430.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science 16*(3), 199–215.

Brezger, A. and S. Lang (2006). Generalized structured additive regression based on bayesian p-splines. *Computational Statistics and Data Analysis 50*(4), 967–991.

Chernoff, H. and S. Zacks (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Annals of Mathematical Statistics 35*, 999–1018.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics 86*(2), 221–241.

Clyde, M. and E. I. George (2004). Model uncertainty. *Statistical Science 19*(1), 81–94.

Cobb, G. W. (1978). The problem of the Nile: conditional solution to a changepoint problem. *Biometrika 65*(2), 243–251.

Cook, R. D. and S. Weisberg (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics 22*(4), 495–508.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics*. published online in advance of print.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General) 147*(2), 278–292.

Denison, D. G. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Chichester: Wiley.

Diggle, P., P. Heagerty, K. Liang, and S. Zeger (2002). *Analysis of longitudinal data* (Second ed.), Volume 25 of *Oxford Statistical Science Series*. Oxford: Oxford University Press.

Doucet, A., N. De Freitas, and N. Gordon (2001). *Sequential Monte Carlo methods in practice*. Statistics for engineering and information science. New York: Springer Verlag.

Fahrmeir, L. and T. Kneib (2010). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford: Oxford University Press.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing 16*(2), 203–213.

Fearnhead, P. and D. Vasileiou (2009). Bayesian analysis of isochores. *Journal of the American Statistical Association 104*(485), 132–141.

Fenske, N., L. Fahrmeir, P. Rzehak, and M. Höhle (2008). Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data. Technical report, Department of Statistics, University of Munich.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*(3), 479–488.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis* (Second ed.). Texts in statistical science. Boca Raton, FL: Chapman & Hall/CRC.

Genest, C. and L.-P. Rivest (2001). On the multivariate probability integral transformation. *Statistics & Probability Letters 53*(4), 391–399.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1998). *Markov chain Monte Carlo in Practice.* Boca Raton, FL: Chapman & Hall.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 243–268.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association 102*(477), 395–378.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test 17*(2), 211–235.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics 19*, 2745–2756.

Held, L. (2008). *Methoden der statistischen Inferenz.* Heidelberg: Spektrum Akademischer Verlag. With the collaboration of Daniel Sabanés Bové.

Held, L., M. Hofmann, M. Höhle, and V. Schmid (2006). A two-component model for counts of infectious diseases. *Biostatistics 7*(3), 422–437.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*(5), 559–570.

Hofmann, M. (2007). *Statistical models for infectious disease surveillance counts.* Ph. D. thesis, Ludwig-Maximilians-Universität München.

Jacob, B., R. Schulz, C. Zorn, I. Lehmann, A. Schötzau, J. Heinrich, G. Sierig, U. Diez, A. Bader, B. Fahlbusch, M. Weiss, L. Jäger, O. Herbarth, B. Schaaf, A. von Berg, H. E. Wichmann, and M. Borte (1999). LISA-Studie: Einfluß von Lebensbedingungen und Verhaltensweisen auf die Entwicklung von Immunsystem und Allergien im Ost-West-Vergleich. *Gesundheitswesen 61*, A100–A101.

Janeway, C. A., S. Carding, B. Jones, J. Murray, P. Portoles, R. Rasmussen, J. Rojo, K. Saizawa, J. West, and K. Bottomly (1988). CD4+ T cells: Specificity and function. *Immunological Reviews 101*(1), 39–80.

Jefferys, W. H. and J. O. Berger (1992). Ockham's razor and Bayesian analysis. *American Scientist 80*(1), 64–72.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures. *Weather & Forecasting 22*(3), 637–650.

Kaslow, R., D. Ostrow, R. Detels, J. Phair, B. Polk, and C. Rinaldo Jr (1987). The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology 126*(2), 310–318.

Kitagawa, G. (1987). Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association 82*(400), 1032–1041.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods Research 33*(2), 188–229.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lamport, L. (1999). *LaTeX. A Document Preparation System. User's Guide and Reference Manual.* Amsterdam: Addison-Wesley Longman.

Lee, C. A., P. B. A. Kernoff, A. N. Phillips, J. Elford, G. Janossy, A. Timms, and M. Bofill (1991). Serial CD4 lymphocyte counts and development of AIDS. *The Lancet 337*(8738), 389–392.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Haerdle and B. Roenz (Eds.), *Compstat 2002 - Proceedings in Computational Statistics*, Heidelberg, pp. 575–580. Physika Verlag.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 44*(1/2), 187–192.

Maguire, B. A., E. S. Pearson, and A. H. A. Wynn (1952). The time intervals between industrial accidents. *Biometrika 39*(1/2), 168–180.

Marshall, E. C. and D. J. Spiegelhalter (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine 22*(10), 1649–1660.

Marshall, E. C. and D. J. Spiegelhalter (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach. *Bayesian Analysis 2*(2), 409–444.

Mosteller, F. and J. W. Tukey (1968). Data analysis, including statistics. In G. Lindzey and E. Aronson (Eds.), *Handbook of Social Psychology*, Volume 2. Addison-Wesley.

Oliver, J. L., P. Carpena, M. Hackenberg, and P. Bernaola-Galván (2004). IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Research 32*, 287–292.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raftery, A. E. and V. E. Akman (1986). Bayesian analysis of a poisson process with a changepoint. *Biometrika 73*(1), 85–89.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Sinharay, S. and H. S. Stern (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference 111*(1), 209–221.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 36*(2), 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 39*(1), 44–47.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test 5*(1), 1–60.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics 35*(6), 2450–2473.

Yao, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics 12*(4), 1437–1447.

I hereby declare that this master thesis has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources or aids have been used in the preparation of this thesis other than those indicated in the thesis itself.

Munich, 20th September 2009

Daniel Sabanés Bové