Overview of Bounded Support Distributions and Methods for Bayesian Treatment of Industrial Data

Kamil Dedecius[1] and Pavel Ettler[2]
[1]*Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic*
[2]*COMPUREG Plzeň, s.r.o., Nádražní 18, 306 34 Plzeň, Czech Republic*

Keywords: Statistical Analysis, Bayesian Analysis, Truncated Distributions, Beta Distribution.

Abstract: Statistical analysis and modelling of various phenomena are well established in nowadays industrial practice. However, the traditional approaches neglecting the true properties of the phenomena still dominate. Among others, this includes also the cases when a variable with bounded range is analyzed using probabilistic distributions with unbounded domain. Since many of those variables nearly fulfill the basic conditions imposed by the chosen distribution, the properties of used statistical models are violated rather rarely. Still, there are numerous cases, when inference with distributions with unbounded domain may lead to absurd conclusions. This paper addresses this issue from the Bayesian viewpoint. It briefly discusses suitable distributions and inferential methods overcoming the emerging computational issues.

1 INTRODUCTION

Modern industrial control systems rely on statistical modelling of various phenomena in the production process, for instance the relevant physical variables, reliability and health of the controlled systems. For this sake, usually traditional approaches providing easy and fast computations are exploited. As an example consider the least-squares based regression or state-space modelling with Kalman filters. Such methods are often explicitly or implicitly based on evaluation of statistical distributions with unbounded support, e.g. the normal distribution. Recognizing the limitations, a number of new methods concerning modelling with bounded support distributions have appeared in the last decade. Their need is obvious:

- Signals occurring in industrial systems are bounded in principle. Limitations start from physical limits of measured signals and margins given by performance of system's actuators, through given ranges of measurement units and, e.g. their A/D converters to limitations given by interpretations of variables within digital computers.
- In many cases, modelling with distributions with unbounded supports can lead to hardly interpretable or even principally impossible values like negative pressure in a hydraulic system, negative rolling force in the rolling mill, negative fuel consumption in engines, reversed direction of current flow in electrical devices, reliability in percents out of interval [0, 100] etc.

This paper focuses on bounded support distributions like the uniform distribution, triangular distributions, beta distribution and its various modifications and truncated normal distribution. We adopt the Bayesian framework, allowing consistent treatment of uncertainty connected with models and estimated values. Since the use of this type of distributions usually calls for approximations (in the Bayesian framework particularly), the expectation-maximization, variational Bayesian inference and Markov chain Monte Carlo methods are discussed as well. The paper ends with an illustrative example of Bayesian beta regression of real rolling mill control data.

2 BAYESIAN INFERENCE

The Bayesian inference denotes a group of statistical methods for estimation of unknown parameters using the Bayes' rule, incorporating new evidence (data) into the prior knowledge in order to obtain posterior knowledge, better reflecting the data-generating reality (system).

Dedecius K. and Ettler P..
Overview of Bounded Support Distributions and Methods for Bayesian Treatment of Industrial Data.
DOI: 10.5220/0004439003800387
In *Proceedings of the 10th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2013), pages 380-387
ISBN: 978-989-8565-70-9

The prior information about the parameter θ, which can be single or multivariate, discrete or continuous, is represented by a probability distribution with a probability density function (pdf) $f(\theta)$. More precisely, it should be written $f(\theta|\alpha)$, where α is a set of parameters of the prior distribution, called *hyperparameters*, to avoid confusion with model parameters. The prior distribution is updated by new observed data x, obeying the *model* (sampling distribution, likelihood) $f(x|\theta)$ via the Bayes' rule

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \quad (1)$$

$$\propto f(x|\theta)f(\theta). \quad (2)$$

While (1) is a full version of the Bayes' rule, in which

$$f(x) = \int f(x|\theta)f(\theta)d\theta$$

is the *marginal* pdf of x, independent of θ and therefore a constant with respect to θ. Its role is to *normalize* the posterior distribution to get a proper distribution with a unit area under the pdf. Notation (2) is a commonly used shorthand for unnormalized pdf.

If the prior distribution of θ is chosen from a class of distributions *conjugate* to the model, then the posterior distribution is of the same type. This particularly appealing fact, providing analytical form of the posterior pdf and allowing dynamic modelling with parameter pdf repeatedly updated and used as prior for the next time step, is connected with the exponential class of distributions.

The Bayesian prediction is provided by the posterior predictive distribution. For a new data point $\tilde{x}$, the predictive distribution given previous data x reads

$$f(\tilde{x}|x) = \int f(\tilde{x}|\theta)f(\theta|x)d\theta.$$

This equation expresses the distribution of a new point *averaged* over the distribution of θ.

Besides the inclusion of prior information and prediction, the Bayesian framework provides many other techniques, most of them consistently built on principles of probability theory. A few examples are systematic model selection and model averaging, hypotheses testing, hierarchical modelling, recursive modelling, distributed parameter estimation etc. Furthermore, being embedded in the dynamic decision making, the Bayesian approach allows to dynamically and adaptively reflect the evolution of reality during modelling.

3 DISTRIBUTIONS WITH BOUNDED SUPPORT

In this section we overview selected *continuous univariate* distributions with bounded support. Their more extensive treatise or dealing with multivariate distributions would exceed the limited extent of the paper. For this reason, we adopt a simplification: only the pdf and the first raw and second central moments are given for each discussed distribution. Also note that the distributions with bounded support may arise either in model or in the prior.

3.1 Uniform Distribution

The uniform distribution of a random variable $X \sim \mathcal{U}(a,b)$ on a compact set $[a,b]$ has the pdf

$$f(x|a,b) = (b-a)^{-1} \quad \text{for } x \in (a,b)$$

and moments

$$\mathbb{E}[X] = \frac{b-a}{2}$$

$$\text{var}\,X = \frac{(b-a)^2}{12}.$$

As a maximum entropy distribution under known support it is suitable merely for cases when no additional knowledge about a random variable is present. In the Bayesian modelling it represents a popular noninformative (vague) prior distribution, however, its use is usually connected with the need of sampling from posterior distribution. An example of uniform pdf on [-0.5, 0.5] is depicted in Fig. 1

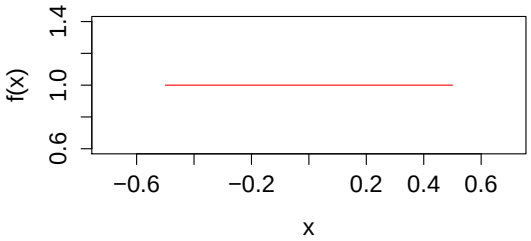

Figure 1: Uniform pdf $\mathcal{U}(-0.5, 0.5)$.

3.2 Triangular Distribution

The triangular distribution on $[a,b]$ with mode $\hat{x}$ of a random variable $X \sim Tri(a,b,\hat{x})$ has the pdf

$$f(x|a,b,\hat{x}) = \begin{cases} 0 & \text{for } x < a, \\ \frac{2(x-a)}{(b-a)(\hat{x}-a)} & \text{for } a \le x \le \hat{x}, \\ \frac{2(b-x)}{(b-a)(b-\hat{x})} & \text{for } \hat{x} < x \le b, \\ 0 & \text{for } x > b. \end{cases} \quad (3)$$

The interesting moments are

$$\mathbb{E}[X] = \frac{a+b+\hat{x}}{3}$$
$$\operatorname{var} X = \frac{a^2+b^2+\hat{x}^2-a(b+\hat{x})-b\hat{x}}{18}.$$

The triangular distributions are suitable for cases when the number of data samples is very limited, preventing reconstruction of possibly more elaborated distribution form. They also arise under certain conditions like the sum of two equivalent uniformly distributed independent random variables, or as a prior distribution of standard deviation under uniformly distributed variance. An example of pdfs with various modes is depicted in Fig. 2.

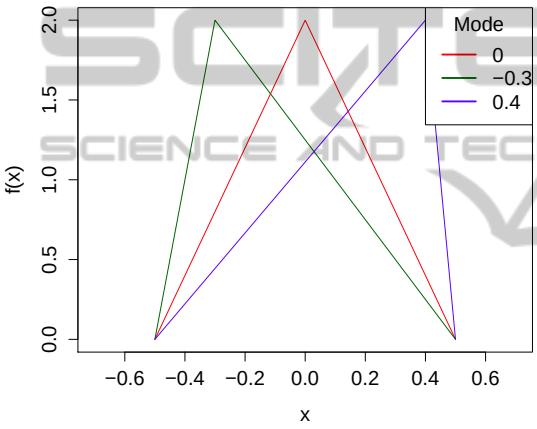

Figure 2: Triangular pdf with $a = -0.5, b = 0.5$ and various modes.

3.3 Beta Distribution

The "basic" beta distribution is a very popular for its flexibility in modelling random variables with bounded range (0,1). Variables with other ranges can be easily transformed by translation and scaling. Its two parameters $\alpha, \beta > 0$ drive the shape of the distribution, allowing for convex and concave shapes, symmetry, left and right skewness and high or low kurtosis and even a flat form of the uniform distribution $\mathcal{U}(0,1)$, see Fig. 3.

The standard pdf of a beta-distributed random variable $X \sim \mathcal{B}(\alpha, \beta)$ has the form

$$f(x|\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is the beta function, $\Gamma(\cdot)$ denotes the gamma function. Under this form, the moments are

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$
$$\operatorname{var} X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

This beta distribution is conjugate to the binomial model $Bi(n,p)$ with parameters $n \in \mathbb{N}$ and $p \in [0,1]$, as the prior for p.

Under several conditions, the parameterization used above may not be suitable. This occurs, e.g., if the random variable X is modelled as a dependent variable given independent regressors. (Ferrari and Cribari-Neto, 2004) propose parametrization with the mean $\mu = \alpha/(\alpha+\beta)$ and precision $\phi = \alpha+\beta$, yielding a beta distribution $\mathcal{B}(\mu\phi, (1-\mu)\phi)$ with pdf

$$f(x|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(1-\mu)\phi)} x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1} \quad (4)$$

with moments

$$\mathbb{E}[X] = \mu$$
$$\operatorname{var} X = \frac{\mu(1-\mu)}{1+\phi}.$$

This form is exploited in beta regression, e.g. (Ferrari and Cribari-Neto, 2004) and (Branscum et al., 2007).

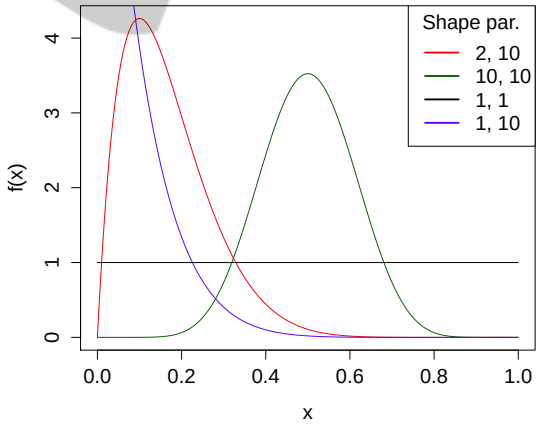

Figure 3: Beta pdf with various shaping parameters α, β.

There exists also a whole class of beta distributions called *generalized beta distributions*, yielding tens of more or less common distributions including χ^2, lognormal, gamma etc. as special cases. The extent of this class is far beyond the scope of this paper.

3.4 Beta-rectangular Distribution

As noted in (Hahn, 2008) the definition of the beta distribution in terms of mean and precision (4) neither considers tail-area events nor greater flexibility

in variance specification. Therefore, (Hahn, 2008) proposed a mixture of beta distribution and a uniform distribution, giving it the name *beta rectangular distribution*. A random variable $X \sim BR(\mu, \phi, \theta)$ has then pdf

$$f(x|\mu,\phi,\theta) = \theta + (1-\theta) f_{\mathcal{B}}(x|\mu,\theta),$$

where μ and ϕ are the mean and precision of a beta component $f_{\mathcal{B}}(\cdot)$ and $\theta \in [0,1]$ is a mixing parameter (weight). Due to the distributions' support, the constant density of the uniform distribution is equivalent directly to θ. The moments of this mixture are straightforwardly

$$\mathbb{E}[X] = \frac{\theta}{2} + (1-\theta)\mu$$

$$\operatorname{var} X = \frac{\mu(1-\mu)}{1+\phi}(1-\theta)\left[1+\theta(1+\phi)\right] + \frac{\theta}{12}(4-2\theta).$$

It is worth to notice that the uniform component is equivalently a beta distribution $\mathcal{B}(1/2, 2)$ and the mixture can be viewed as a beta mixture with one component fixed. The beta-uniform mixture was recently proposed to improve robustness of beta regression to outliers, (Bayes et al., 2012). Some examples of the beta-rectangular distribution are in Fig. 4.

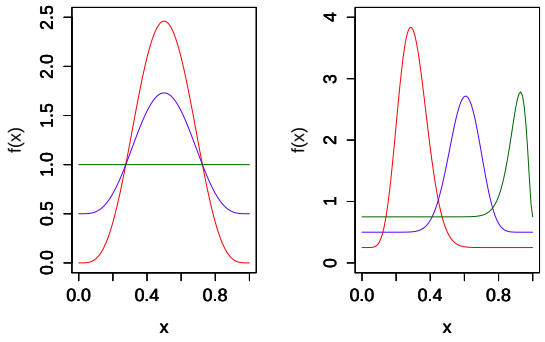

Figure 4: Various forms of beta-rectangular pdf. Left: beta component with parameters $\mu = 0.5, \phi = 10$ and $\theta = 0$ (red), $\theta = 0.5$ (blue), $\theta = 1$ (green). The special cases in red and green correspond to pure beta and uniform distributions, respectively. Right: parameters $[\theta, \mu, \phi]$: [0.25, 0.3, 30] (red), [0.5, 0.6, 30] (blue) and [0.75, 0.9, 30] (green).

3.5 Inflated Beta Distributions

Another approach to the issue of tail-area events are the so-called *inflated beta distributions*. Similarly to the beta-rectangular distribution (beta-uniform mixture), the inflated beta distributions are themselves mixtures. The pdf of a zero-inflated (or one-inflated) beta distribution can be written in the form

$$f_c(x|\theta,\mu,\phi) = \theta \mathbf{1}_c(x) + (1-\theta) f_{\mathcal{B}}(x|\mu,\phi), \quad (5)$$

where $c = 0$ or $c = 1$ for zero-inflated and one-inflated distribution, respectively; $f_{\mathcal{B}}(x|\mu,\phi)$ is a beta pdf (4) and $\mathbf{1}_c(x)$ is an indicator function equal to one if $x = c$ and zero otherwise. The interest in bounded-support distributions implies that the zero-and-one-inflated beta distribution should be expressed similarly to (5) with two weights for the cases $c = 0$ and $c = 1$, with the complement to 1 reserved for the beta component, similarly to (Rigby and Stasinopoulos, 2005).

Another expression of zero-and-one-inflated beta was proposed in (Ospina and Ferrari, 2010) as a mixture of beta and Bernoulli distributions with pdf

$$f_{\{0,1\}}(x|\theta,p,\mu,\phi) = \theta f_{Ber}(x;p) + (1-\theta) f_{\mathcal{B}}(x;\mu,\phi), \quad (6)$$

where $f_{Ber}(\cdot)$ is a Bernoulli component with parameter $p \in [0,1]$ and $f_{\mathcal{B}}(\cdot)$ is again a beta component[1]. An example is depicted in Fig. 5. The first and second moments are then

$$\mathbb{E}[X] = \theta p + (1-\theta)\mu$$

$$\operatorname{var} X = \theta p(1-p) + (1-\theta)\frac{\mu(1-\mu)}{1+\phi} + \theta(1-\theta)(p-\mu)^2.$$

The inflated beta distributions have been used in the

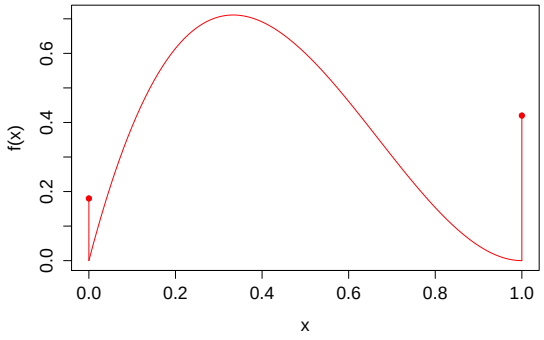

Figure 5: Zero-and-one-inflated beta pdf with parameters $\theta = 0.6, p = 0.3, \mu = 0.4$ and $\phi = 5$.

non-Bayesian generalized linear models (GLM), e.g. (Rigby and Stasinopoulos, 2005).

3.6 Truncated Normal Distribution

From the class of truncated distributions we choose the normal one. The truncated normal distribution follows from restricting the normal distribution with a

[1]The discrete Bernoulli component in (6) is also understood as a pdf, that is, a Radon-Nikodým derivative of of the distribution with respect to a dominating *counting* measure.

mean μ and variance σ^2 to a bounded interval, either from one or both sides, by truncation, i.e., by cutting the tails out of interest and subsequent renormalization. A random variable obeying the (two-sided) truncated normal distribution, $X \sim t\mathcal{N}(\mu,\sigma^2,a,b)$ has the pdf

$$f(x|\mu,\sigma^2,a,b) = \frac{\sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)},$$

where $\phi(\cdot)$ denotes the pdf of the standard normal distribution $\mathcal{N}(0,1)$ and $\Phi(\cdot)$ used for renormalization is its distribution function. The moments are

$$\mathbb{E}[X] = \mu + \frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\sigma$$

$$\mathrm{var}X = \sigma^2\left[1 + \frac{\frac{a-\mu}{\sigma}\phi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma}\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right] - \sigma^2\left(\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right)^2$$

Four examples of truncated normal distribution on interval [-0.5, 0.5] are depicted in Fig. 6.

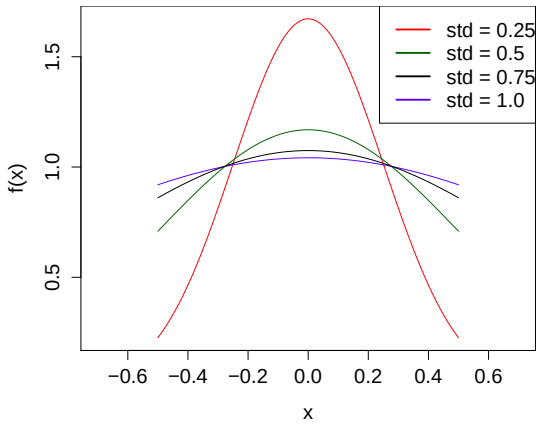

Figure 6: Truncated normal pdf with $a = -0.5, b = 0.5$, mean $\mu = 0$ and various standard deviations.

The truncated normal distribution is popular in parameter estimation, however, it was almost disregarded in the industrial practice. Another distribution with a great potential is the truncated Student's t distribution, whose heavier tails provide more robust statistical computations. The main issues of truncated distributions is that analytical computations are rarely possible and a sort of rather demanding Monte Carlo method (like Gibbs sampling) is usually unavoidable.

4 COMPUTATIONAL ISSUES

In many cases, the Bayesian inference of the posterior distribution of parameters θ given data is analytically intractable. The difficulty lies in the normalizing constant of the posterior distribution, which often contains special functions (c.f. the beta distribution, normalized by a beta function). Among the most popular ways around the problems are stochastic simulations, particularly Markov Chain Monte Carlo (MCMC) and the expectation-maximization (EM) algorithm. In addition to them, we discuss also the variational Bayesian inference (Jordan, 1999) that has become a popular analytical approach often providing similar accuracy as the Gibbs sampler, but at a much greater speed.

4.1 Expectation Maximization

The Expectation-Maximization (EM) algorithm for two-stage iterative finding of maximum likelihood or maximum a posteriori (MAP) estimates of latent variables of probabilistic models was proposed by (Dempster et al., 1977). Similarly to the notation introduced earlier, we consider a model of observed variables X governed by a set of parameters θ. Suppose, that direct optimization of the likelihood $f(X|\theta)$ is difficult and, under existence of discrete or continuous hidden variable Z, optimization of $f(X,Z|\theta)$ is much easier. Recall, that the marginalization rule yields

$$f(X|\theta) = \int f(X,Z|\theta)dZ,$$

providing a way to the desired likelihood. It is straightforward to show that the existence of a distribution $q(Z)$ allows to write

$$\log f(X|\theta) = \mathcal{L}(q,\theta) + \mathcal{D}(q||f), \qquad (7)$$

where

$$\mathcal{D}(q||f) = -\int q(Z)\log\frac{f(Z|X,\theta)}{q(Z)}dZ \qquad (8)$$

is the Kullback-Leibler divergence of q and f (Kullback and Leibler, 1951). It is a premetric, i.e. a nonnegative functional satisfying $\mathcal{D}(q||f) = 0$ if $q = f$ almost everywhere. The term

$$\mathcal{L}(q,\theta) = \int q(Z)\log\frac{f(X,Z|\theta)}{q(Z)}dZ \qquad (9)$$

is a lower bound of (7), which reaches maximum if

$$q(Z) = f(Z|X,\theta).$$

Starting from a crude estimate θ^*, the two steps of the EM algorithm are

1. Expectation (E-step) – the lower bound $\mathcal{L}(q,\theta^*)$ is maximized with respect to $q(Z)$ with the current value θ^* being fixed. The aim is to get $q(Z)$ as close to $f(Z|X,\theta^*)$ as possible, that is to minimize the Kullback-Leibler divergence in (7). By substitution $q(Z) = f(Z|X,\theta^*)$ into (9),

$$\begin{aligned}\mathcal{L}(q,\theta) &= \mathbb{E}_{Z|X,\theta^*}[\log f(X,Z|\theta)] \\ &\quad - \mathbb{E}_{Z|X,\theta^*}[\log f(Z|X,\theta^*)] \\ &= Q(\theta,\theta^*) + \text{const.}, \end{aligned} \tag{10}$$

   where the constant independent of θ is the entropy of q.

2. Maximization (M-step) – the pdf $q(Z)$ is fixed and $\mathcal{L}(q,\theta)$ is maximized with respect to θ in order to obtain a new value θ'. This increases the lower bound and consequently the log likelihood in (7). Since the distribution $q(Z)$ is fixed, the Kullback-Leibler divergence to $f(Z|X,\theta')$ is positive. Using (10),

$$\theta' = \arg\max_{\theta} Q(\theta,\theta^*).$$

Then $\theta^* \leftarrow \theta'$ and the algorithm is repeated until convergence.

The EM algorithm and its variants like GEM (generalized EM) or ECM (expectation conditional maximization) are particularly popular in estimation of finite mixtures. Here, the latent variable Z denotes the component from which the available data originate.

4.2 Variational Bayes

The variational Bayesian (VB) inference, rooted in the field of calculus of variations, serves for analytic approximation of the posterior pdf of parameters and potentially other latent variables (Jaakkola and Jordan, 2000). Let us denote $Z = (Z_1,\ldots,Z_n)$ as the set comprising both parameters and latent variables. The goal is to find analytically tractable approximation $q(Z)$ of $f(Z|X)$. Similarly to EM decomposition (7), we may write

$$\log f(X) = \mathcal{L}(q) + \mathcal{D}(q||f), \tag{11}$$

where the analogues of (8) and (9) are

$$\mathcal{D}(q||f) = -\int q(Z)\log\frac{f(Z|X)}{q(Z)}dZ$$

$$\mathcal{L}(q) = \int q(Z)\log\frac{f(Z|X)}{q(Z)}dZ.$$

Unlike in the EM algorithm, the elements of Z are factorized into M *independent* factors $Z_i, i = 1,\ldots,M$, such that

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i).$$

This, put back into (11) yields

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_i q_i(Z_i)\left[\log f(X,Z) - \sum_i \log q_i(Z_i)\right]dZ \\ &= \int q_j(Z_j)\mathbb{E}_{i\neq j}[\log f(X,Z)]dZ \\ &\quad - \mathbb{E}_j[\log q_j(Z_j)] + \text{const.}\end{aligned}$$

where

$$\mathbb{E}_{i\neq j}[\log f(X,Z)] = \int \log f(X,Z)\prod_{i\neq j} q_i(Z_i)dZ_i.$$

This directly yields the VB-optimal factors

$$\log q_j^*(Z_j) = \mathbb{E}_{i\neq j}[\log f(X,Z)] + \text{const.}$$

The additive constant changes to multiplicative in exponentiation, providing the solution

$$q_j^*(Z_j) \propto \exp\{\mathbb{E}_{i\neq j}[\log f(X,Z)]\}.$$

The resulting algorithm is very similar to the expectation-maximization, but unlike it, VB computes the posterior distributions of all parameters. The expectations are taken with respect to variables not in the current factor, which, in turn, are recomputed in the same way. The algorithm is guaranteed to converge and, under convexity of the lower bound, to the global maximum (Boyd and Vandenberghe, 2004).

It is necessary to stress that the variational Bayesian method provides analytic approximations of the posterior distribution of parameters and latent variables. The sacrifice is their factorized treatment (11), neglecting the dependency properties carried by the true joint posterior pdfs. An alternative *expectation propagation* algorithm (Minka, 2001) overcomes this issue by exploiting reversed order of pdfs in the Kullback-Leibler divergence in (11). The price is elevated level of computational difficulties.

A recent example of the VB algorithm used in conjunction with bounded variables is presented in (Ma and Leijon, 2011). It provides a method for VB estimation of beta mixture models. An interesting part of the paper is approximate analytic solution of otherwise analytically intractable integrals emerging from special (gamma or beta) functions in the beta distribution. This reveals the pervasive computational problems connected even with the very standard distributions with bounded support.

4.3 Simulation from Posterior

The industrial practice often deals with complicated models, for which the inference is neither analytically nor approximately (in the EM and VB sense) tractable. This issue is yet emphasized when distributions with bounded support are used. The form

of resulting analytically unreachable posteriors need to be evaluated by simulations, exploiting a (usually big) set of draws to represent the distributions. In high-dimensional problems, the Markov chain Monte Carlo (MCMC) methods dominate this field. The idea of Markov chain simulation is to simulate a random walk in the space of unknown (multivariate) parameter θ. The random walk converges to a stationary distribution close to the target posterior $f(\theta|x)$ (Gelman et al., 2003). Two popular MCMC methods, the Metropolis-Hastings (Metropolis et al., 1953) and Gibbs algorithms (Geman and Geman, 1984), have become standards in Bayesian modelling.

Metropolis-Hastings Algorithm: first draws a starting point θ accomplishing $f(\theta|x) > 0$ from some suitable distribution. Then, it recursively exploits a Markov transition kernel (proposal distribution) $q(\theta'|\theta)$ in the following way:

1. Sample a candidate point θ' from $q(\theta'|\theta)$.
2. Calculate
$$r = \min\left(1, \frac{f(\theta'|x)q(\theta|\theta')}{f(\theta|x)q(\theta'|\theta)}\right). \qquad (12)$$
3. Move to θ' with probability r, else stay at θ.

The choice of the Markov transition kernel is generally uneasy: the main requirements, besides easy sampling and computing of r, is a reasonable distance travelled in the parameter space with each transition and high acceptance rate. For example, a normal kernel with high variance leads to a low acceptance rate as the proposed samples often lay in regions with small pdf value. On the other hand, a very low variance produces high acceptance rate in regions with high pdf, but it takes a lot of iterations until the proposed samples explore more distant regions (which is called slow *mixing*).

Gibbs Sampling: is a special case of the Metropolis-Hastings algorithm. It divides θ into m parts $\theta = (\theta_1, \ldots, \theta_m)$. Each iteration cycles through these components, drawing each subset conditional on the value of all the others (Gelman et al., 2003). First, an ordering of components is chosen at random and each θ'_i is sampled from the conditional distribution $f(\theta'_i|\theta'_1, \ldots, \theta'_{i-1}, \theta'_{i+1}, \ldots, \theta_m)$. The Markov transition kernel can be shown to have a special form allowing jumps only of those components of θ' that match the previous θ. Under this condition, the value of r in equation (12) is always 1, thus every jump is accepted. This suppresses slow mixing of the chain.

In both the Metropolis-Hastings and Gibbs algorithms, the initial values can be chosen either randomly or using initial points from some crude approximation (e.g. via EM). This is also the reason for discarding a subset of the first iterations (usually several thousands) called *burn-in*, that is likely far from the target distribution. The drawback of both algorithms is the correlation of samples, significantly influencing the mixing and other properties of simulations. Also fine tuning of MCMC methods is a tedious time-consuming task, requiring careful prior and post-hoc analyses to ensure that the simulated values of θ are close to the target distribution of $\theta|y$.

5 EXAMPLE OF APPLICATION

As an illustrative example, we estimate the Bayesian beta regression model (e.g. (Ferrari and Cribari-Neto, 2004) and (Branscum et al., 2007)) on a 20 data points from a rolling mill, depicted in Fig. 7. The horizontal axis represents discrete time, the vertical axis describes the control in 0.01%. We fitted the model (4) using a logit link function as follows

$$y_i|\mu_i, \phi \sim B(\mu_i\phi, (1-\mu_i)\phi)$$
$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, 20.$$

corresponding with the reparameterized beta distribution (4). The coefficients (β_0, β_1) together with precision ϕ were estimated as independent,

$$f(\beta_0, \beta_1, \phi) = f(\beta_0, \beta_1) f(\phi)$$

with β_0 and β_1 being normal and ϕ gamma distributed.

The model was estimated in GNU R interfacing with OpenBUGS through the BRugs package. The chain length was 50 000 samples with initial 4000 samples serving for burn-in.

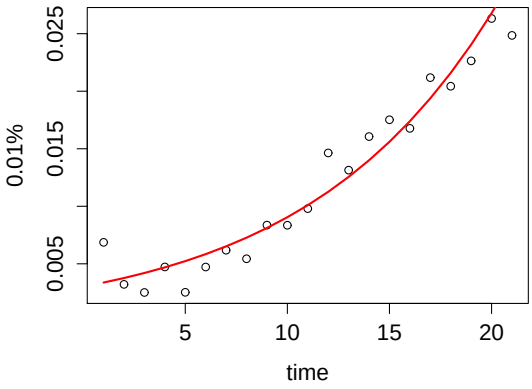

Figure 7: Time × control in 0.01%. Points are true measurements, red line is interpolation obtained by beta regression.

Results of estimation of regression coefficients are given in Table 1. The mean values of posterior distributions are $\hat{\beta}_0 = -5.799$ and $\hat{\beta}_1 = 0.11$, with the corresponding 95% credibility intervals (defined as highest density intervals) being $[-7.762, -3.827]$ and $[-1.853, 2.075]$ for β_0 and β_1, respectively. One of the rules of thumb recommends that the simulation should be run until the Monte Carlo error for each parameter of interest falls below 5% of the sample standard deviation. Table 1 shows that the simulation reached less than 0.6% for both coefficients.

The posterior distributions of β_0 and β_1 are depicted in Fig. 8 as histograms of Monte Carlo samples together with kernel density estimates (in red).

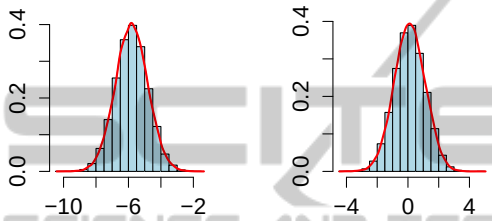

Figure 8: Bayesian beta regression – posterior distributions of β_0 (left) and β_1 (right). Histograms depict relative frequency of MCMC samples, red lines are respective kernel density estimates.

Table 1: Results of MCMC estimation of beta regression model. $\tilde{\boldsymbol{x}}_{2.5}$, $\tilde{\boldsymbol{x}}_{50}$ and $\tilde{\boldsymbol{x}}_{97.5}$ denote 2.5%, 50% and 97.5% quantiles of posterior distributions.

	β_0	β_1
mean	-5.799	0.110
st. dev.	1.001	1.005
MC error	5.486e-3	5.177e-3
$\tilde{\boldsymbol{x}}_{50}$	-5.802	0.113
$\tilde{\boldsymbol{x}}_{2.5}$	-7.762	-1.853
$\tilde{\boldsymbol{x}}_{97.5}$	-3.827	2.075
MC error/stdev	0.548%	0.515%

For comparison, the *betareg* package was used for beta regression in frequentist statistical framework (Ferrari and Cribari-Neto, 2004). The model had the same structure, the link function was identically the logit. Coefficients estimates were $\hat{\beta}_0 = -5.866$ and $\hat{\beta}_1 = 0.115$, respectively, model precision was 2578.

ACKNOWLEDGEMENTS

The research project is supported by the grant MŠMT 7D12004 (E!7262 ProDisMon).

REFERENCES

Bayes, C. L., Bazán, J. L., and García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, 7(4):841–866.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, 49(3):287–301.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741.

Hahn, E. (2008). Mixture densities for project management activity times: A robust approach to PERT. *European Journal of Operational Research*, 188(2):450–459.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jordan, M. I. (1999). An introduction to variational methods for graphical models. In *Machine Learning*, pages 183–233.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Ma, Z. and Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2160–2173.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ospina, R. and Ferrari, S. (2010). Inflated beta distributions. *Statistical Papers*, 51:111–126. 10.1007/s00362-008-0125-4.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.