Graphs For Machine Learning
Quasi-Bayesian Survival Analysis

Adèle Gillier
adele.gillier@ens-paris-saclay.fr

Victor Sanh
victor.sanh@ens-paris-saclay.fr

Instructor: Benjamin Guedj
benjamin.guedj@inria.fr

Abstract

This project aims to explore an application of graphs in machine learning: survival analysis. We investigate the use of quasi-Bayesian methods to learn and predict survival. We shows that the quasi-Bayesian framework is an interesting framework for doing inference applied to survival analysis.
First we will introduce quasi-Bayesian inference (in opposition to Bayesian inference) and recall some important notions and notations of survival analysis. Then, we present two simple models (exponential model and Cox piece-wise constant hazard model) for analyzing survival times and write down the equations for performing the quasi-Bayesian inference. Lastly, we perform numerical experiments on simple simulated data to validate our approach, and illustrate the main differences, strengths and weaknesses of Bayesian VS quasi-Bayesian inference.

1 Bayesian VS Quasi-Bayesian framework

1.1 Bayesian learning

The Bayesian paradigm is based on specifying a probability model for the observed data D, given an unknown parameter vector β which leads to a likelihood function we will denote $L(D|\beta)$. We assume that the parameter β is a random variable and we can formulate a prior distribution on β: $\Pi(\beta)$. Intuitively, this prior is the knowledge an expert can formulate on the distribution of β before seeing the observations. Inference on β is then computed with Bayes rule:

$$\begin{aligned} \Pi(\beta|D) &= \frac{L(D|\beta)\Pi(\beta)}{\int_B L(D|\beta)\Pi(\beta)d\beta} \\ \Pi(\beta|D) &\propto L(D|\beta)\Pi(\beta) \end{aligned} \tag{1}$$

where $\Pi(\beta|D)$ denotes the posterior distribution, and B is the parameter space in which β lies. We will call the denominator $\int_B L(D|\beta)\Pi(\beta)d\beta$ the *marginalization distribution of the data* or *the normalization constant*.

Often, the normalization constant does not have a closed analytical expression, and hence, we know the posterior distribution only up to a constant. However, we can still numerically sample from the posterior distribution using approximate inference techniques such as Monte Carlo Markov Chains. For instance, Gibbs sampling and Metropolis-Hastings algorithm are commonly used.

Once we have estimated the posterior distribution, an interesting task is prediction. If we denote $\Pi(y|D)$ the posterior prediction distribution given the data, we can write:

$$\Pi(y|D) = \int_B f(y|\beta)\Pi(\beta|D)d\beta \tag{2}$$

MVA 2017/2018, Graphs in Machine Learning project.

where $f(y|\beta)$ denotes the sampling distribution of y. This formula expresses the uncertainty that we have on β by marginalizing the distribution of $y|\beta$ by the posterior distribution of $\beta|D$. Hence, sampling from the posterior predictive distribution will often lead to a wider range of values than taking the predictive distribution obtained by using the single estimate $\hat{\beta}$ from the posterior distribution (maximum a posteriori or mean a posteriori are commonly used).

1.2 Quasi-Bayesian learning

The Quasi-Bayesian paradigm is slightly different than the Bayesian paradigm. It relies on the idea of replacing the likelihood by a term related to the prediction error (also called risk).

Let's suppose that we have of a set of observations $(Y_1, ...Y_n)$. We want to build a reasonable prediction $\hat{Y}$ of Y ($\hat{Y}$ can be thought as $\hat{Y} = \hat{\psi}(X)$ where $\hat{\psi}$ is a prediction function to determine, and X is a feature vector). We denote l a loss function. We define the risk R and the empirical risk R_n: $R = \mathbb{E}\big[l(\hat{Y}, Y)\big]$ and $R_n = \frac{1}{n}\sum_{i=1}^{n} l(\hat{Y}_i, Y_i)$.

Moreover, we have a set of predictors $\mathbb{F} = \{\hat{\psi}\}$, and a probability measure π over $\mathbb{F}$. For convenience, $\mathbb{F}$ can be though as a set of parameters that characterize the predictors (equivalent of B in the Bayesian framework). We define $\alpha > 0$ a temperature parameter.

Hence, the quasi-Bayesian framework relies on the following definition of the quasi-posterior $\hat{\rho}$:

$$\hat{\rho}(\cdot) \propto \exp(-\alpha R_n(\cdot))\pi(\cdot) \tag{3}$$

Formula 3 is the equivalent of 1 in the Bayesian framework. We have replaced the likelihood term by a term related to the empirical risk: $\exp(-\alpha R_n(\cdot))$. The term *quasi* refers to the fact that the distribution obtained after inference is not a true posterior.

Like in the Bayesian framework, it is often difficult to obtain a closed analytic form for the normalization constant. Hence, we will use MCMC techniques to sample from the quasi-posterior distribution and compute a single predictor from the quasi-posterior. We can usually use the Maximum A Quasi-Posterior (equivalent of MAP) $\hat{\psi} \in \arg\max_{\psi} \hat{\rho}(\psi)$ or the Mean A Quasi-Posteriori $\hat{\psi} = \int_{\mathbb{F}} \psi \hat{\rho}(d\psi)$.

It is important to note that we have dropped the dependence on α (the temperature parameter) in $\hat{\rho_\alpha}$ and $\hat{\psi_\alpha}$. When $\alpha \to 0$, the quasi-posterior looks like the prior (observations have limited impact), and when $\alpha \to +\infty$, observations have more influence on the quasi-posterior and the predictors will tend to only minimize the empirical risk (overfitting of the training examples). To follow the quasi-Bayesian philosophy, we can also consider α as a random variable, define a prior over it and integrate it into the quasi-Bayesian inference. We have now to estimate both the predictors and the temperature control. This point will be further discussed during numerical experimentation.

2 Survival analysis

2.1 General review and notations

Survival analysis is the analysis of the expected duration of time until one event happens, such as an accident, death, or failure in mechanical systems [6]. The goal is to understand the variables and their impact onto a survival function, to be able to predict the survival time of individuals and/or determine the important variables. The treatment history for a patient can be modeled by a graph, nodes represent treatments over time and edges model the survival probability. Survival analysis is used, for instance, on medical trial to estimate the impact of a medical treatment on patients.

We study the random variable T representing the survival time of an individual. We note $f(t)$ its probability density function. The probability of an individual's surviving till time t is given by the survivor function $S(t) = \mathbb{P}(T > t)$. An important variable to study in survival analysis is the instantaneous risk of failure at time t, noted $h(t)$, also called the *hazard function*. It is defined as the probability that the accident happens at time t, knowing that the individual as survived until t:

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T \leq t + dt)}{S(t)dt} = \frac{f(t)}{S(t)} \tag{4}$$

As $1 - S(t)$ is the cumulative distribution function, we have $-S'(t) = f(t)$, and then $S(t) = \exp\Big(-\int_0^t h(u)du\Big)$.

Censoring

Usually, experiments are made on a population under a certain time range. The survival times of the individuals are only observed if the accident happens during the time range. An observation is said to be right-censored at C, if we only observe survival times happening before the time C. For the other individuals we only know that they survived until time C.

We introduce the following notations:

- $Y_i = \min(T_i, C)$ is the observed survival times of the individual i right-censored at C.
- $\Delta_i = 1$ if $T_i \leq C$ and $\Delta_i = 0$ otherwise.

2.2 A Bayesian framework for survival analysis

Let's consider n individuals that we observe until the right-censorship C. The survival times $(T_i)_{i=1..n}$ of the individuals are independent and identically distributed (i.i.d.) with density f and survival function S. We observed n pairs of variables $(Y_i, \Delta_i)_{i=1..n}$. The probability density of the variable Y_i is: $\mathbb{P}(Y_i = t) = f(Y_i)^{\Delta_i} S(Y_i)^{1-\Delta_i}$. Indeed, if the individual is right-censored, i.e. $\Delta_i = 0$, then $Y_i = C$ and its probability is $\mathbb{P}(T_i > C) = S(C) = S(Y_i)$. The likelihood, knowing the parameters of the model β, can be written as following:

$$L(Y, \Delta|\beta) = \prod_{i=1}^{n} f(Y_i|\beta)^{\Delta_i} S(Y_i|\beta)^{1-\Delta_i} \tag{5}$$

Then, by assuming some prior on β, we can determine the a posteriori probability distribution, and estimate the parameter β.

2.3 An important particular model: the Cox hazard regression model

To determine the impact of individual's description features X on the survival time, we often use the Cox hazard regression model, also called the Proportional hazard model. This model makes the assumption that the logarithm of the instantaneous risk depends linearly on the data X. We can write the hazard function, or instantaneous risk:

$$h(t|X) = h_0(t) \exp(< \beta, X >) \tag{6}$$

with $h_0(t)$ the baseline risk function, and β a vector of regression coefficients. The survival function for patient i can then be written: $S(Y_i) = \exp\big[-\exp(X_i'\beta) \int_0^{Y_i} h_0(u)du\big]$. The goal is then to estimate the regressors β.

3 Quasi-Bayesian survival analysis

In this section, we present two models we have constructed for survival analysis and set up a quasi-Bayesian framework for inference.

3.1 Exponential model - Quasi-Bayesian resolution

We consider an exponential model: we suppose that T_i the survival time for patient i follows an exponential distribution of unknown parameter λ_i for i in $\{1, ..., n\}$, where n denotes the number of patients considered in the experiment. We introduce C, the right-censorship time, and observe Y_i: $Y_i = C$ if T_i is right-censored, or $Y_i = T_i$ if not. To introduce the regression model, we will suppose that each patient i is described by a vector $X_i \in \mathbb{R}^p$ and we set $\lambda_i = \exp(X_i'\beta)$, where β is a vector of p coefficients.

Using this parameterized model, we can build a prediction $\hat{Y}_i$ for each patient i. We will use the mean as predictor: $\hat{Y}_i = \mathbb{E}(Y_i)$. By definition, we have $Y_i = \mathbb{1}_{\{T_i<C\}} T_i + \mathbb{1}_{\{T_i \geq C\}} C$.

Using the known parameterized distribution of T_i and integration by parts, we have:

$$\begin{aligned}\hat{Y}_i &= \mathbb{E}(\mathbb{1}_{\{T_i<C\}}T_i) + C\mathbb{P}(T_i \geq C) \\ &= \frac{1-\exp(-\lambda_i C)}{\lambda_i} \\ &= \frac{1-\exp(-e^{X_i'\beta}C)}{e^{X_i'\beta}}\end{aligned} \tag{7}$$

As we are comparing time values, it seem reasonable to consider the mean squared loss. The empirical risk is then given by:

$$\begin{aligned}R_n(\hat{Y}, Y) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\exp(-\lambda_i C)}{\lambda_i} - Y_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\exp(-e^{X_i'\beta}C)}{e^{X_i'\beta}} - Y_i\right)^2\end{aligned} \tag{8}$$

The goal of inference is thus outputting a reasonable estimate of the parameter vector β given the data. Before doing quasi-Bayesian inference, we need to define prior on β: $\pi(\beta)$. We usually consider a uniform prior on β (it means that we know nothing relevant about β) assuming that the coefficients of β are bounded. If we have more prior knowledge, we could also use a normal or a gamma prior. The quasi-Bayesian inference equation is then given by:

$$\hat{\rho}(\beta|D) \propto \exp\left(-\frac{\alpha}{n}\sum_{i=1}^{n}\left(\frac{1-\exp(-e^{X_i'\beta}C)}{e^{X_i'\beta}} - Y_i\right)^2\right)\pi(\beta) \tag{9}$$

As previously said, α could also be integrated into the prior and the quasi-posterior. In our experiments, we used the mean as a quasi-posterior estimation: $\hat{\beta} = \mathbb{E}_{\hat{\rho}}[\beta]$.

3.2 Cox piecewise constant hazard model - Quasi-Bayesian resolution

In this subsection, we propose a second model which is an adaptation of the Cox piecewise constant hazard model. We will reuse the same notations introduced in the last subsection.

The time of experiment is partitioned into a fixed set of J time intervals: $\{I_j = [s_{j-1}, s_j[; s_0 = 0; j = 1..J\}$. Besides, we assume that $s_{J-1} < C < s_J$. For convenience, we will assume that $s_J = +\infty$. On each time partition, the Cox piecewise constant hazard model assumes that the baseline risk function is constant: $\forall 1 \leq j \leq J, \forall s_{j-1} \leq t < s_j, h_0(t) = \lambda_j$.

The relation between the density function and the hazard function is:

$$f(t) = \exp\Big(-\int_0^t h(u)du\Big)h(t) \tag{10}$$

Besides, we introduce the covariates for the regression model as follows (X represents the p features of a patient):

$$h(t) = h_0(t)\exp(X'\beta) \tag{11}$$

We still assume that β is a constant vector of size p. Hence, we can write an explicit formula for $\int_0^t h_0(u)du$ and the density function can be expressed as follows:

$$\forall s_0 \leq t \leq s_J, f(t) = \prod_{j=1}^{J}\left[(\lambda_j e^{x'\beta})^{\delta_j} \cdot \exp\Big(-\delta_j\big\{\lambda_j(t-s_{j-1}) + \sum_{i=1}^{j-1}\lambda_i(s_j - s_{j-1})\big\}e^{x'\beta}\Big)\right] \tag{12}$$

where $\delta_j = \mathbb{1}_{s_{j-1}\leq t \leq s_j}$.

We assume that T takes f as a density function. We still use the mean prediction $\hat{Y} = \mathbb{E}(Y)$ and using integration by parts, we can write an explicit formula for the predictor:

$$\begin{aligned}
\hat{Y} = \sum_{j=1}^{J-1} \Bigg[& \frac{\exp\Big(e^{x'\beta}\big(\lambda_j s_{j-1} - \sum_{i=1}^{j-1}\lambda_i(s_i - s_{i-1})\big)\Big)}{\lambda_j e^{x'\beta}} \\
& \times \Big(\exp(-\lambda_j e^{x'\beta} s_{j-1})(1+\lambda_j e^{x'\beta} s_{j-1}) - \exp(-\lambda_j e^{x'\beta} s_j)(1+\lambda_j e^{x'\beta} s_j)\Big)\Bigg] \\
+ & \Bigg[\frac{\exp\Big(e^{x'\beta}\big(\lambda_J s_{J-1} - \sum_{i=1}^{J-1}\lambda_i(s_i - s_{i-1})\big)\Big)}{\lambda_J e^{x'\beta}} \\
& \times \Big(\exp(-\lambda_J e^{x'\beta} s_{J-1})(1+\lambda_J e^{x'\beta} s_{J-1}) - \exp(-\lambda_J e^{x'\beta} C)(1+\lambda_J e^{x'\beta} C)\Big)\Bigg] \\
+ & \exp\Big(e^{x'\beta}\big(\lambda_J s_{J-1} - \sum_{i=1}^{J-1}\lambda_i(s_i - s_{i-1})\big)\Big)\Big(\exp(-\lambda_J e^{x'\beta} C) - \exp(-\lambda_J e^{x'\beta} s_J)\Big)
\end{aligned} \tag{13}$$

The next steps are the same as previously presented: $\hat{\rho}(\beta) \propto \exp\big(-\alpha R_n(\beta)\big)\pi(\beta)$ and the final estimator for β is: $\hat{\beta} = \mathbb{E}_{\hat{\rho}}[\beta]$.

4 Numerical experimentation

We simulate a dataset where we have several information about each individual, we denote by X_i the information about the individual i, such as age, medical information, etc... We fix a value for β and simulate the survival time for each individual following the exponential model of parameter $\lambda_i = \exp(X_i^T \beta)$, and censored at the time $t = 100$. This corresponds to the exponential survival model presented at the section 3.1. By doing this way, we know the ground truth β and are able to simply compare models. The metric used in the following sections to compare models is the difference between the parameter estimated by the simulation $\hat{\beta}_i$ and the ground truth β_i, that we know since we simulate our database.

To estimate the parameter β we draw samples from the quasi posterior law $\hat{\rho}(\beta) \propto \pi(\beta)\exp(-\alpha R_n(\beta, X))$ using Monte Carlo Markov Chain sampling methods. We implement the estimator using the *Python* package for sampling: *pymc*. The code is available at `https://github.com/VictorSanh/Quasi-Bayesian_Survival_Analysis`.

4.1 The number of training examples

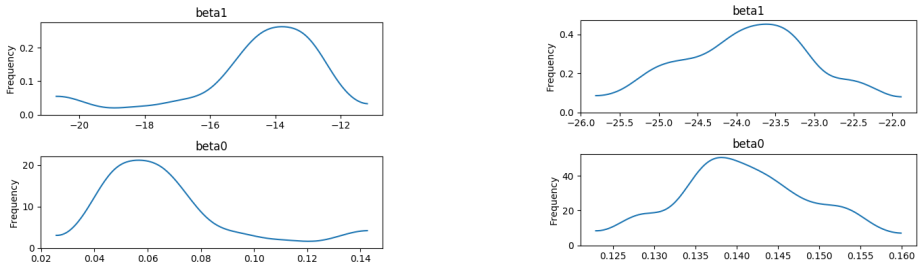
First, we study the influence of the number of training examples in the database. All things being equal, we only make the number of training examples varies. Globally, the more there are training examples, the more accurate the estimation of β is (refer to Table 1). Nonetheless, we observe a saturation effect, for more than 100 training examples adding more training examples does not significantly improve the quality of the estimation. This is a strength of this type of inference: we can obtain a reasonable estimation with a small number of observations. Figure 1 displays the quasi-posterior distribution for 5 and 250 training examples. The quasi-posterior distribution for 5 observations seems particularly unlikely and unreasonable (compared to the one obtained for 250 training examples).

4.2 The temperature parameter α

The temperature parameter α is a term of trade-off between the prior knowledge and the data. As we said, the higher the parameter α is, the more influent the data are over the priors. α have to be chosen regarding the prior knowledge about β, and the amount of available data. If we have a good prior knowledge about β we should put a small value on α to give importance on the prior. Furthermore if there is not enough data available, give a small value to α will prevent the model from overfitting. On

Number of Training Examples	$\beta_0 - MeanQP(\beta_0)$	$\beta_1 - MeanQP(\beta_1)$
5	0.0912	-11.0215
25	-0.0554	4.2482
50	-0.0123	5.5768
100	-0.1214	1.6209
250	0.0099	-1.2420

Table 1: Different size of training dataset. Quasi-posteriori estimations are improving with the number of training examples.



(a) 5 training examples

(b) 250 training examples

Figure 1: Different size of training dataset. Quasi-posterior distributions are improving with the number of training examples.

the other hand if there are very few information about the values of β and a large uniform prior is put on, then α should be high to give more importance to the data.

In the example we study, we have a lot of available data, we observe then better results when α is higher. Table 2 shows the gap between the estimated parameter $\hat{\beta}$ and the ground truth for several values of α. If $\alpha \to \infty$ then the model is equivalent to minimizing the empirical risk. We could also consider α as a random variable, put a prior on it and draw samples according to the quasi-posterior $\pi(\beta, \alpha) \exp(-\alpha R_n(.))$.

Value of α	$\beta_0 - MeanQP(\beta_0)$	$\beta_1 - MeanQP(\beta_1)$
100	-0.0017	-0.2478
10	-0.0039	0.6376
1	-0.0102	1.3560
0.01	0.0108	-0.6733
$U([0, 0.1])$	0.0297	-0.4784

Table 2: Mean A Quasi Posteriori for several number of training examples

4.3 The choice of the priors over the parameter β

Tables 3 and 4 show the gap between the estimated parameter $\hat{\beta}$ and the ground truth. We do not observe significant difference when the prior distribution over β_0 or β_1 change. However we have chosen distribution that are all in the same range, which is the range of the parameter. It seems that there is no significant change when the prior is "good enough".

Prior over β_0	$\beta_0 - MeanQP(\beta_0)$	$\beta_1 - MeanQP(\beta_1)$
$\Gamma(\alpha = 1, \beta = 10)$	0.0043	-0.4743
$\Gamma(\alpha = 1, \beta = 4)$	0.0100	-1.1312
$\mathcal{N}(0.1, 5)$	0.0087	-0.1011
$\mathcal{N}(0.15, 5)$	0.0109	-1.2317
$U([-1, 1])$	0.0089	-0.9939
$U([0, 1])$	0.0052	-0.5875

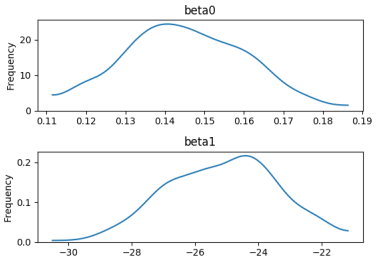
Table 3: Mean A Quasi-Posteriori for several priors over β_0 (the ground truth β_0 is equal to 0.15)

Prior over β_1	$\beta_0 - MeanQP(\beta_0)$	$\beta_1 - MeanQP(\beta_1)$
$\mathcal{N}(-25, 10)$	0.0074	-0.1517
$\mathcal{N}(-25, 20)$	0.0074	-0.2629
$U([-30, -20])$	0.0050	0.1155
$U([-40, -20])$	0.0026	-0.1003
$U([-60, 0])$	0.0050	-0.1114

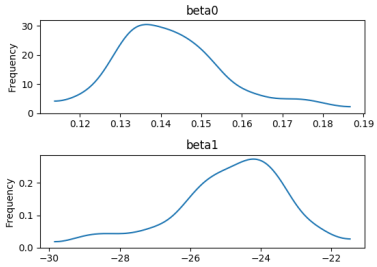
Table 4: Mean A Quasi-Posteriori for several priors over β_1 (the ground truth β_1 is equal to -25)

Figures 2 and 3 show the quasi-posterior distribution of the parameters β_0 and β_1 for different prior distributions. We can see that the shape of the quasi-posterior distribution is really different from the shape of the prior. (when giving a uniform prior, the quasi-posterior is more similar to a normal distribution than to a uniform distribution).

These plots also suggest that we could use the quasi-posterior as the prior of a future inference on the same class of data. It enables us to form a a rough idea of a relevant prior knowledge to include in the inference (using historical data or a simpler model).

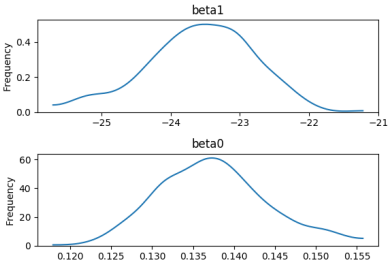


(a) Prior over $\beta_1 \sim \mathcal{N}(-25, 20)$

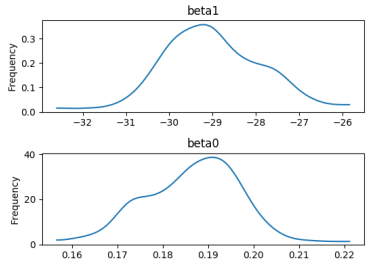


(b) Prior over $\beta_1 \sim U([-40, -20])$

Figure 2: Different quasi posteriori distributions for different priors over β_1. Prior over β_0 is always the same: $\mathcal{N}(0.1, 5)$



(a) Prior over $\beta_0 \sim \Gamma(\alpha = 1, \beta = 4)$



(b) Prior over $\beta_0 \sim U([0, 1])$

Figure 3: Different quasi-posterior distributions for different priors over β_0. Prior over β_1 is always the same: $\mathcal{N}(-25, 5)$

4.4 Bayesian experimentation

We use the same generated dataset to estimate the parameter β using a Bayesian method to compare the two methods. The log likelihood of the exponential model is given by:

$$l(y_i, \Delta_i, \lambda_i) = \sum_i \Delta_i \log(X_i^T \beta) - \sum_i y_i X_i^T \beta = \log(p_\Delta(\beta|X))$$

We choose the mean posterior to estimate the parameter β. To do so, we draw values of $\hat{\beta}$ following the distribution $\propto \pi(\beta) p_\Delta(\beta|X)$ with $\pi(\beta)$ a prior over β and $p_\Delta(\beta|X)$ the likelihood. The final estimator of β if then the mean of the drawn values.

The simulations of the Bayesian estimator were done in *Python* using the package *pymc* to sample, as in quasi-Bayesian simulations. With similar priors we obtain less better results. Indeed we obtain values of $\hat{\beta}_0$ around 0.06 and values of $\hat{\beta}_1$ around -18, whereas the true values are $\beta_0 = 0.15$ and $\beta_1 = -25$, for different priors. Even though those estimators are more distant to the ground truth than the ones given by the quasi-Bayesian method, there are still in a correct range. On this simple survival model quasi-Bayesian method seems to give better result, however more dataset and models have to be simulate to find out if this result can be generalized.

5 Discussion: Bayesian VS Quasi-Bayesian framework

We will try to give a high-level discussion of the strengths and weaknesses of the quasi-Bayesian inference framework for survival analysis and compare it to the Bayesian inference framework.

In the quasi-Bayesian framework, the observations $(Y_1, ..., Y_n)$ are deterministic and no likelihood is attached to our approach, hence the term quasi-posterior. In this framework, we do not need any statistical model on the distribution of the data, the only way observations play a role in the inference is through the risk term. We can for instance use a "black-box" algorithm to predict Y, where we do not know the form of the distribution but we do know the risk as a function of several parameters.

On the other hand, quasi-Bayesian inference can be tricky because of the temperature parameter which can greatly influence the quality of the posterior prediction. This temperature parameter can be difficult to fine tune depending on the influence of past observation we want to impose. However, as discussed in the section 4.2, the temperature parameter can be chosen to leverage the importance of data and prior knowledge. A well chosen parameter can give an information of how precise is the prior knowledge, or try to avoid overfitting by putting more importance on prior when few data are available. The Bayesian framework has no simple equivalent of this temperature parameter. This can be seen as a strength of the quasi-Bayesian approach, but also as a weakness because of the difficulty of choosing a reasonable value.

There remain a challenge in both Bayesian and quasi-Bayesian framework: the choice of the priors. As suggested by the experimentation, both the posterior (or quasi-posterior) distribution are influenced by the prior imposed, and as long as the prior is close enough to the true value, the posterior and quasi-posterior will be reasonably accurate.

6 Future work and conclusion

In this study, we have only used simulated data, on which we know the ground truth. A natural extension of this work is to use the model developed here on real data. However, on these data, we do not know the ground truth for regression parameters β. The natural way we could use to compare a Bayesian and quasi-Bayesian approach is to fix a test set and compute the mean squared errors on the test set for each inference. We use the same approach if we want to compare a classical machine learning algorithm applied to survival analysis to a quasi-Bayesian approach. Note that there is no easy equivalent of the Bayesian inference's Bayes factor in the quasi-Bayesian framework.

To conclude, we have demonstrated that the quasi-Bayesian framework is an interesting framework to tackle survival analysis. In particular, we have developed two different models (an exponential model, and a Cox piece-wise constant hazard model) and implemented the quasi-Bayesian and Bayesian inference on simulated data for the first model. We have studied the influence on the estimation on the regression parameters of several parts such as the size of the training set, the priors and the temperature parameter. The first results suggests the quasi-Bayesian approach is better at taking into account the data than Bayesian approach, but further simulation on other dataset or model have to be done to elaborate and precise the conclusions of this study.

References

[1] Françis Bach and Guillaume Obozinski. Introduction to probabilistic graphical models. 2017.

[2] Olivier Bouaziz. Analyse de survie: le modèle de cox.

[3] David R. Cox. *Regression Models and Life-Tables*, pages 527–541. Springer New York, New York, NY, 1992.

[4] Benjamin Guedj. *Aggregation of estimators and classifiers : theory and methods*. Theses, Université Pierre et Marie Curie - Paris VI, December 2013.

[5] Benjamin Guedj. Quasi-bayesian learning. 2017.

[6] Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian survival analysis*. Wiley Online Library, 2005.

[7] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *CoRR*, abs/1708.04649, 2017.