Stockholms universitet

Comparison of Bayesian and ML predictive distributions

Gustav Thorzén

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2025:9
Matematisk statistik
Juni 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Mathematical Statistics
Stockholm University
Bachelor Thesis **2025:9**
http://www.math.su.se

Comparison of Bayesian and ML predictive distributions

Gustav Thorzén*

June 2025

Abstract

Maximum likelihood estimation (MLE) is one of, if not, the most well known and used method of point estimation. Predictive distributions created from substitution of the true unknown parameters with the MLE are shown to be inferior in terms of average Kullback-Leibler (KL) divergence for independent and identically distributed exponential and unknown mean normal random variables. A novel technique for calculating densities of Bayesian predictive distributions, also known as posterior predictive distributions, is provided and illustrated for exponential and normal random variables.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
Supervisor: Ola Hössjer, Johannes Heiny.

Table of Contents

1. Introduction

A common problem in statistics is that of forecasting future events. What probabilities should be assigned to future events given the data we have seen? One of the most common ways is to assume the future events/data come from some, a priori, known distributional class with fixed unknown parameters, and heuristically estimate the parameters from the data seen so far. This is the category of point estimators, to create a predictive distribution by substituting the unknown true parameters with their respective point estimators. The quality of these methods is typically judged by the repeated sampling properties of the point estimators.

This thesis compares predictive distribution created using one of the most well know method of creating point estimators, that of maximum likelihood estimators, with a much less known Bayesian method. Instead of using point estimators in the place of the true parameters, the parameters are treated as random variables and the predictive distribution is obtained by integrating the unknown parameters with the respect to a posterior distribution. This is referred to in this thesis as Bayesian predictive distributions, but sometimes referred to as posterior predictive distributions. Bayesian predictive distributions, unlike point estimators based ones, have a direct interpretation of representing one's belief in future outcome, given the data seen so far, under the assumption of a known distributional class with fixed unknown parameters.

In this first section, the introduction, we present the contents of the following sections along with the acknowledgements. The second section presents the notation, prerequisite theory, definitions that will be used, and a technique to calculate the Bayesian predictive distribution from the normalization constant of the posterior distribution. It also contains a large sample size asymptotic distribution independent result, for the average KL divergence, provided by one of my supervisors. The third section provides theoretical motivation for the superiority of the Bayesian predictive distribution over the ML predictive distribution, in the case of an exponential model. In this section we also provide a derivation of it and the maximum likelihood estimator. The fourth section provides theoretical motivation for of superiority of the Bayesian predictive distribution over the ML predictive distribution, in the case of a normal model with unknown mean and known variance. In this section we also provide a derivation of the Bayesian predictive distribution for a normal model with unknown mean and unknown variance. The fifth section concludes the thesis summarizing the results.

Acknowledgements

Special thanks goes to my supervisors Ola Hössjer & Johannes Heiny who not only allowed me to go ahead with the ideas I pitched, but also greatly helped me to untangle what notation to use to minimize confusion. They also helped me to find sources which enabled me to obtain key results analytically, and more.

2. General Theory

For a random variable X with parameter λ, with λ_0 being the unknown true value, and sample $x_{1:n} = (x_1, \dots x_n)$ of size n, the following notation will be used through out the thesis:

$f(x|\lambda)$ = Density function
$F(x|\lambda)$ = Cumulative distribution function
$\pi_0(\lambda)$ = Prior distribution of λ
$\pi(\lambda|x_{1:n})$ = Posterior distribution of λ
$g(x|x_{1:n})$ = Bayesian predictive density function
$G(x|x_{1:n})$ = Bayesian predictive distribution function
$I(\lambda)$ = Fisher information

2.1. Maximum Likelihood

The method of maximum likelihood estimation (MLE) creates, for an unknown fixed parameter λ, an estimator $\hat{\lambda}_{ML}$ as the solution to the equation

$$0 = \sum_{k=1}^{n} \left(\frac{d}{d\lambda} \log(f(x_k|\hat{\lambda}_{ML})) \right)$$

where n is the sample size of an i.i.d. sample, so that x_k are observations of independent identically distributed random variables X_k with density function $f(x|\lambda)$ for som parameter λ. In other words $\hat{\lambda}_{ML}$ is the point estimator maximizing the chance to obtain the given sample. By substituting the true unknown parameter λ_0 with $\hat{\lambda}_{ML}$ we obtain the density $f(x|\hat{\lambda}_{ML})$ as the density function of the ML predictive distribution.

Since the MLE is the value of λ with maximizes the sum of the empirical log-likelihoods

$$\sum_{k=1}^{n} \left(\log(f(x_k|\hat{\lambda}_{ML})) \right)$$

by definition, it follows that it minimizes

$$n^{-1} \sum_{k=1}^{n} \left(\log\left(\frac{1}{f(x_k|\hat{\lambda}_{ML})} \right) \right)$$

and therefore also minimizes

$$n^{-1} \sum_{k=1}^{n} \left(\log\left(\frac{f(x_k|\lambda_0)}{f(x_k|\hat{\lambda}_{ML})} \right) \right)$$

which means it asymptotically as $n \to \infty$ it minimizes

$$\int_{x_-}^{x_+} \left(f(x|\lambda_0) \log\left(\frac{f(x|\lambda_0)}{f(x|\hat{\lambda}_{ML})} \right) \right) dx$$

which can, from the definition of KL divergence in section 2.6, be seen to be the KL divergence between the true density and the ML predictive density. The is MLE therefore asymptotically optimal in this regard.

2.2. Fisher Information

The Fisher information of a, univariate, parameter λ is defined for a random variable X with density function $f(x|\lambda)$, supported on (x_-, x_+), where $-\infty \le x_- < x_+ \le \infty$, as follows:

$$I(\lambda) = \int_{x_-}^{x_+} f(x|\lambda_0)\left(\frac{d}{d\lambda}\log(f(x|\lambda))\right)^2 dx$$

and when Fisher's regularity conditions hold it can also be calculated through

$$-\int_{x_-}^{x_+} f(x|\lambda_0)\left(\frac{d^2}{d\lambda^2}\log(f(x|\lambda))\right)dx$$

assuming $\log(f(x|\lambda))$ is twice differentiable with respect to λ.

2.3. Jeffreys's Prior

Within a Bayesian framework, the Jeffreys's prior is a well known prior distribution, defined as the square root of the determinant of the Fisher information. In other words $\pi_0(\lambda) = |I(\lambda)|^{1/2}$. It is noteworthy that Jefferys's prior is an equivariant prior, meaning that if we change the parametrization of the distribution of the random variable, and recalculate the posterior, it will be the same had we simply reparameterized the posterior. This means that Jefferys's prior represents the same prior information independent of parametrization.

2.4. Bayesian predictive distribution

The predictive distribution of a random variable corresponds to our believed distribution given our model assumptions and the data we have seen so far. The Bayesian predictive distribution is fundamentally different from point estimator based predictive distributions in that it corresponds to a weighted average. More specifically, if

$$f(x_{1:n}|\lambda) = \prod_{k=1}^{n} f(x_k|\lambda)$$

is the likelihood of the data, and with $\pi(\lambda|x_{1:n})$ as the posterior, defined as:

$$\pi(\lambda|x_{1:n}) = \frac{f(x_{1:n})\pi_0(\lambda)}{\int_{\lambda_-}^{\lambda_+} f(x_{1:n})\pi_0(\lambda)d\lambda}$$

then the Bayesian predictive density function is defined as:

$$g(x|x_{1:n}) = \int_{\lambda_-}^{\lambda_+} f(x|\lambda)\pi(\lambda|x_{1:n})d\lambda,$$

where $-\infty \le \lambda_- < \lambda_+ \le \infty$ are the lower and upper bounds for the parameter. The Bayesian predictive distribution function, similar to the cumulative distribution function, is defined as:

$$G(x|x_{1:n}) = \int_{x_-}^{x} g(t|x_{1:n})dt,$$

with $G(x_-|x_{1:n}) = 0$, and $G(x_+|x_{1:n}) = 1$.

If the dataset $x_{1:n}$ has a sufficient statistics of the form $s_b = \sum_{k=1}^{n} b(x_k)$ for some $b(x)$, there exist another way to compute the density of the Bayesian predictive distribution. Calculate as a function of the sufficient statistics:

$$p(n, s_b) = \int_{\lambda_-}^{\lambda_+} f(x_{1:n}|\lambda)\pi_0(\lambda)d\lambda,$$

where n is the sample size. We then get the density of the Bayesian predictive distribution as

$$g(x|n, s_b) = \frac{p(n+1, s_b + h(x))}{p(n, s_b)},$$

conveniently parameterized in terms of the sufficient statistics. This will be the way we calculate the Bayesian predictive densities later.

A proof of this is easily seen using the definition of the predictive density. Start with the definition of the predictive density:

$$g(x|x_{1:n}) = \int_{\lambda_-}^{\lambda_+} f(x|\lambda)\pi(\lambda|x_{1:n})d\lambda$$

then expand the posterior distribution into

$$\pi(\lambda|x_{1:n}) = \frac{f(x_{1:n}|\lambda)\pi_0(\lambda)}{p(n, s_b)}$$

where $p(n, s_b)$ can be moved out of the integral, because it is independent of λ, to obtain

$$g(x|x_{1:n}) = \frac{\int_{\lambda_-}^{\lambda_+} f(x|\lambda)f(x_{1:n}|\lambda)\pi_0(\lambda)d\lambda}{p(n, s_b)}$$

where the numerator is rewritable to $p(n+1, s_b + h(x))$ and finally get

$$g(x|x_{1:n}) = \frac{p(n+1, s_b + h(x))}{p(n, s_b)}$$

which is better expressed as $g(x|n, s_b)$.

2.5. Bernstein-von Mises Theorem

The Bernstein-von Mises Theorem connects the MLE with the Bayesian posterior distribution. Specificly it states that the posterior distribution converges, asymptotically in total variation distance, to a normal distribution centered at the maximum likelihood estimator, and with a variance based on the Fisher information.

$$||\pi(\cdot|x_{1:n}) - N(\hat{\lambda}_{ML}, (nI(\lambda_0))^{-1})||_{TV} \to 0$$

This means that the posterior distribution asymptotically converges to a distribution of a single point $P(\lambda = \lambda_0) = 1$, and therefore the Bayesian predictive density, $g(x|x_{1:n})$, asymptotically converges to the true density, $f(x|\lambda_0)$.

2.6. KL Divergences

The Kullback-Leibler (KL) divergence between two density function $h_0(x)$ and $h_1(x)$ is defined as:

$$KL(h_0(x), h_1(x)) = \int_{x_-}^{x_+} h_0(x) \log\left(\frac{h_0(x)}{h_1(x)}\right)dx$$

and it can in information theory be interpreted as the average additional information needed to encode the outcome from a random variable with density $h_0(x)$ when assuming it instead have the density $h_1(x)$.

Using the KL divergence between the true and the ML or Bayesian predictive distribution, we define for a specific sample

$$DKL_{ML} = KL\left(f(x|\lambda_0), f(x|\hat{\lambda}_{ML})\right)$$

$$DKL_B = KL\Big(f(x|\lambda_0), g(x|x_{1:n})\Big)$$

where the first letter D emphasizes that both KL divergences are functions of a sample and therefore have a sampling distribution in the context of repeated sampling. For repeated sampling we define

$$EKL_{ML}(n) = \int_{x_{1:n-}}^{x_{1:n+}} DKL_{ML}\, f(x_{1:n}|\lambda_0) dx_{1:n}$$

$$EKL_B(n) = \int_{x_-}^{x_+} DKL_B\, f(x_{1:n}|\lambda_0) dx_{1:n}$$

as functions of the sample size n. They correspond to the expected KL divergence between the true distribution and the ML and Bayesian predictive distributions respectively.

Both DKL_{ML} and DKL_B can however be split into two parts, with the differential entropy as a common term, and a separate cross entropy term. The differential entropy is defined as

$$\int_{x_-}^{x_+} f(x|\lambda_0) \log\Big(f(x|\lambda_0)\Big) dx$$

and the cross entropy terms as

$$-\int_{x_-}^{x_+} f(x|\lambda_0) \log\Big(f(x|\hat{\lambda}_{ML})\Big) dx$$

for the ML predictive distribution and

$$-\int_{x_-}^{x_+} f(x|\lambda_{0)} \log\Big(g(x|x_{1:n})\Big) dx$$

for the Bayesian predictive distribution respectively.

A distribution independent asymptotic result for both $EKL_{ML}(n)$ and $EKL_B(n)$, courtesy of my supervisor Ola, states that for a parameter vector $\lambda = (\lambda_1, \ldots, \lambda_p)$:

$$EKL_{ML}(n) = \frac{p}{2n} + o(1/n)$$

$$EKL_B(n) = \frac{p}{2n} + o(1/n)$$

where $o(1/n) \to 0$ as $n \to \infty$ and $p = dim(\lambda)$, which for univariate λ means $p = 1$. Proof is provided in the appendix.

3. Exponential Model

The density and distribution functions using the λ rate parametrization are:

$$f(x|\lambda) = \lambda \exp(-\lambda x)$$
$$F(x|\lambda) = 1 - \exp(-\lambda x)$$

where $x_- = 0$ and $x_+ = \infty$.

3.1. Maximum Likelihood

The MLE $\hat{\lambda}_{ML}$ is ns_1^{-1}. To see this, we look for the solution of the likelihood equation

$$0 = \sum_{k=1}^{n} \left(\hat{\lambda}_{ML}^{-1} - x_k \right)$$

which in terms of the sufficient statistic s_1, were $h(x) = x^1$, is

$$\hat{\lambda}_{ML} = ns_1^{-1}$$

3.2. Fisher Information and Jeffreys's Prior

The Fisher information is given by $I(\lambda) = \lambda^{-2}$. To see this we calculate

$$I(\lambda) = \int_0^{\infty} \frac{\lambda \exp(-\lambda x)}{\lambda^2} \, dx$$

were the denominator can be moved out of the integral since it does not depend on x, to trivially obtain the solution λ^{-2} knowing the density integrates to 1.

From the above Fisher information we get Jeffreys's prior $\pi_0(\lambda) = I(\lambda)^{1/2} = \lambda^{-1}$.

3.3. Posterior Distribution and Normalizer

Using Jeffreys's prior as our prior distribution, $\pi_0(\lambda)$, the posterior becomes a gamma distribution with $\alpha = n$ and $\beta = s_1$. To see this we calculate the unnormalized posterior

$$f(x_{1:n}|\lambda)\pi_0(\lambda) = \lambda^{-1} \prod_{k=1}^{n} \lambda \exp(-\lambda x_k) = \lambda^{n-1} \exp(-s_1 \lambda)$$

whose integral, the posterior normalizer $p(n, s_1)$, is $\Gamma(n) s_1^{-n}$.

Proof: Since the density $f_Z(z|\alpha, \beta)$ of a gamma distributed random variable Z with parameters α and β integrates to 1, we get

$$1 = \int_0^{\infty} \beta^n z^{\alpha-1} \exp(-\beta z) \Gamma(\alpha)^{-1} dz.$$

We multiply both sides with $\Gamma(\alpha)\beta^{-\alpha}$, and substitute $z = \lambda$, $\alpha = n$, $\beta = s_1$, to get

$$\Gamma(n) s_1^{-n} = \int_0^{\infty} \lambda^{n-1} \exp(-s_1 \lambda) d\lambda$$

and by dividing $f(x_{1:n})\pi_0(\lambda)$ by $p(n, s_1)$, the previous gamma distribution can be obtained as the posterior distribution.

3.4. Bayesian Predictive Distribution

Knowing the posterior normalizer $p(n, s_1) = \Gamma(n) s_1^{-n}$, we have $p(n+1, s_1 + x) = \Gamma(n+1)(s_1 + x)^{-(n+1)}$. We divide them to get the desnisty function of the predictive distribution:

$$g(x|n,s_1)=\frac{p(n+1,s_1+x)}{p(n,s_1)}=\frac{\Gamma(n+1)(s_1+x)^{-(n+1)}}{\Gamma(n)s_1^{-(n)}}$$

a special case of the generalized Pareto distribution sometimes called a Lomax distribution, but rarely written in this form since it can be simplified to something much more intuitive.

Since sample sizes cannot be non-integers we can use $k!=\Gamma(k+1)$, and rewrite s_1^{-n} as $s_1 s_1^{-(n+1)}$, to get

$$g(x|n,s_1)=\frac{n(s_1+x)^{-(n+1)}}{s_1 s_1^{-(n+1)}}$$

which can be further simplified into

$$g(x|n,s_1)=\left(\frac{n}{s_1}\right)\left(1+\frac{x}{s_1}\right)^{-(n+1)}$$

or in terms of the maximum likelihood estimator $\hat{\lambda}_{ML}$

$$g(x|n,\hat{\lambda}_{ML})=\hat{\lambda}_{ML}\left(1+\frac{\hat{\lambda}_{ML}x}{n}\right)^{-(n+1)}$$

where it can easily be seen that $g(x|n,\hat{\lambda}_{ML})$ converges to the true density $f(x|\lambda_0)$ as $n\longrightarrow\infty$ since

$$\lim_{n\to\infty}\hat{\lambda}_{ML}\left(1+\frac{\hat{\lambda}_{ML}x}{n}\right)^{-(n+1)}=\lambda_0\exp(-\lambda_0 x)=f(x|\lambda_0)$$

using the well known theorem

$$\lim_{n\to\infty}\left(1+\frac{z}{n}\right)^{(n+1)}=\exp(z).$$

and applying $\exp(-z)=\exp(z)^{-1}$ with the substitution $z=\hat{\lambda}_{ML}x$.

The cumulative distribution function, parameterized by the sufficient statistic and MLE respectively are

$$G(x|n,s_1)=1-\left(1+\frac{x}{s_1}\right)^{-n}$$

$$G(x|n,\hat{\lambda}_{ML})=1-\left(1+\frac{\hat{\lambda}_{ML}x}{n}\right)^{-n},$$

which is easily seen by taking their derivatives to obtain the above density, and noting it is 0 for $x_-=0$ and 1 for $x_+=\infty$.

3.5. Resulting KL Divergences

Now knowing the MLE and the density of the Bayesian predictive distribution, we can compare them in terms of KL divergence, but first we calculate the common differential entropy term of DKL_{ML} and DKL_B:

$$\int_0^\infty f(x|\lambda_0)\log\left(f(x|\lambda_0)\right)dx=\int_0^\infty f(x|\lambda_0)\left(\log(\lambda_0)-\lambda_0 x\right)dx=\log(\lambda_0)\int_0^\infty f(x|\lambda_0)dx-\lambda_0\int_0^\infty f(x|\lambda_0)x\,dx=\log(\lambda_0)-1$$

having separated the two integrals and moved out the terms independent of x, the final step uses the fact the density integrates to 1, and the expectation value of λ_0^{-1}.

With the common differential term known, the calculation of DKL_{ML} becomes

$$DKL_{ML} = \log(\lambda_0) - 1 - \int_0^\infty f(x|\lambda_{0)} \log\left(\hat{\lambda}_{ML} \exp\left(-\hat{\lambda}_{ML} x\right)\right) dx$$

which can be simplified to

$$DKL_{ML} = -\log\left(\frac{\hat{\lambda}_{ML}}{\lambda_0}\right) - 1 + \int_0^\infty f(x|\lambda_0)\left(\hat{\lambda}_{ML} x\right) dx$$

and futher into

$$DKL_{ML} = -\log\left(\frac{\hat{\lambda}_{ML}}{\lambda_0}\right) - 1 + \frac{\hat{\lambda}_{ML}}{\lambda_0}$$

and by shifting back into sufficient statistics

$$DKL_{ML} = \log\left(\frac{\lambda_0 s_1}{n}\right) - 1 + \frac{n}{\lambda_0 s_1}$$

and through a substitution $t_1 = \lambda_0 s_1$ get

$$DKL_{ML} = \log\left(\frac{t_1}{n}\right) - 1 + \frac{n}{t_1}$$

where t_1 for an unrealized sample would follow a Gamma distribution with $\alpha = n$ and $\beta = 1$.

With the common differential entropy term known, the calculation of DKL_B becomes

$$DKL_B = \log(\lambda_0) - 1 - \int_0^\infty f(x|\lambda_0) \log\left(\hat{\lambda}_{ML}\left(1 + \frac{\hat{\lambda}_{ML} x}{n}\right)^{-(n+1)}\right) dx$$

which can be simplified into

$$DKL_B = -\log\left(\frac{\hat{\lambda}_{ML}}{\lambda_0}\right) - 1 + (n+1) \int_0^\infty f(x|\lambda_0) \log\left(1 + \frac{\hat{\lambda}_{ML} x}{n}\right) dx$$

and by switching to sufficient statistics, substituting $z = \lambda_0 x$ and using a different normalized sufficient statistic $t_1 = \lambda_0 s_1$

$$DKL_B = \log\left(\frac{t_1}{n}\right) - 1 + (n+1) \int_0^\infty f(z|1) \log\left(1 + \frac{z}{t_1}\right) dz$$

As a next step we expand the density function $f(z|1)$, an exponential random variable with unit rate parameter, so that the Bayesian KL divergence becomes

$$DKL_B = \log\left(\frac{t_1}{n}\right) - 1 + (n+1) \int_0^\infty \exp(-z) \log\left(1 + \frac{z}{t_1}\right) dz.$$

After solving the integral, using a proof in the appendix , the Bayesian KL divergence, finaly becomes

$$DKL_B = \log\left(\frac{t_1}{n}\right) - 1 + (n+1) \exp(t_1) \Gamma(0, t_1)$$

were $\Gamma(0, z) = \int_t^\infty exp(-z)/z \, dz$ is the upper incomplete gamma function, see Abramowitz & Stegun (1964) page 262 (6.5.15) for more details.

3.6. Average KL Divergences

Knowing the formula for DKL_{ML} and DKL_B, written in terms of n and t_1, for a specific sample, we would like to know their repeated sampling behavior, specifically their average.

In repeated sampling t_1 will be a gamma random variable with $\alpha = n$ and $\beta = 1$. Using here $f(t_1)$ as its density function, we have

$$EKL_{ML} = \int_0^\infty DKL_{ML}(t_1) f(t_1) dt_1$$

$$EKL_B = \int_0^\infty DKL_B(t_1) f(t_1) dt_1$$

from the definitions in section 2.

Starting with EKL_{ML} we immediately expand DKL_{ML} and get

$$EKL_{ML} = \int_0^\infty f(t_1) \left(\log(t_1) - \log(n) - 1 + \frac{n}{t_1} \right) dt_1$$

where $f(t_1)$ is the density. We can split the integral and substitute the logarithmic and inverse moments, see the proof in the appendix, to get

$$EKL_{ML} = \psi(n) - \log(n) - 1 + \frac{n\Gamma(n-1)}{\Gamma(n)}$$

Knowing that sample sizes can only take integer values lets us simplify the last expression into

$$EKL_{ML} = \psi(n) - \log(n) + \frac{1}{n-1}$$

with $\psi(n) = \Gamma'_n(n)/\Gamma(n)$ representing the digamma function, see Abramowitz & Stegun (1964) page 258 (6.3.1) for more details.

Going next for EKL_B we again expand DKL_B and get

$$EKL_B = \int_0^\infty f(t_1) \left(\log(t_1) - \log(n) - 1 + (n+1) \exp(t_1) \Gamma(0, t_1) \right) dt_1$$

where we again split the integral and substitute the moments, see the proof in the appendix, to get

$$EKL_{PB} = \psi(n) - \log(n) + \frac{1}{n}$$

a very similar but smaller result then that of EKL_{ML}. Specifically $EKL_{ML} - EKL_B = 1/(n(n-1))$.

Using the inequality (2.2) $1/(2x) \le \log(x) - \psi(x) \le 1/(x)$, which can be found in Alzer (1997) page 374, we can rewrite it as $0 \le \psi(x) - \log(x) + 1/x \le 1/(2x)$ to obtain a further result that $EKL_B \le 1/(2n)$ and see that it outperforms the distribution independent asymptotic result of $1/(2n)$.

4. Normal Model

The density function using the mean and standard deviation parametrization is

$$f(x|\mu,\sigma) = \left(2\pi\sigma^2 \exp\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-1/2}$$

but it will be more convenient using a different less common parametrization

$$f(x|\mu,v) = \left(2\pi v \exp\left(\frac{(x-\mu)^2}{v}\right)\right)^{-1/2}$$

with mean μ and variance $v = \sigma^2$.

4.1. Maximum Likelihood, Fisher Information, and KL divergence

The maximum likelihood estimator $\hat{\mu}_{ML}$ and $\hat{v}_{ML}$ are the well known $s_1 n^{-1}$ and $\frac{s_2}{n} - \left(\frac{s_1}{n}\right)^2$ respectively. Notably the marginal distribution of $\hat{\mu}_{ML}$ is another normal with mean μ_0 and variance v_0/n.

The Fisher information matrix for the normal model is

$$\begin{pmatrix} v^{-1} & 0 \\ 0 & (2v^2)^{-1} \end{pmatrix}$$

with the top left element for the mean and the bottom right element for the variance.

The KL divergence between to normal densities is

$$\int_{-\infty}^{\infty} f(x|\mu_0,v_0)\log\left(\frac{f(x|\mu_0,v_0)}{f(x|\mu_1,v_1)}\right)dx = \frac{(\mu_0-\mu_1)^2}{2v_1} + \frac{1}{2}\left(\frac{v_0}{v_1} - 1 - \log\left(\frac{v_0}{v_1}\right)\right)$$

where the first one has mean μ_0 and variance v_0 and the second has mean μ_1 and variance v_1.

4.2. Unknown Mean with Known Variance

From the square root of the Fisher information of the mean we have Jeffreys's prior as $\pi_0(\mu) = v^{(-1/2)}$ and the posterior normalization constant as

$$p(n,s_1,s_2) = \int_{-\infty}^{\infty} f(x_{1:n}|\mu,v)\pi_0(\mu)d\mu = \int_{-\infty}^{\infty}\left((2\pi)^n v^{(n+1)}\exp\left(\frac{s_2 - 2s_1\mu + n\mu^2}{v}\right)\right)^{(-1/2)} d\mu$$

which we can rewrite into

$$p(n,s_1,s_2) = \int_{-\infty}^{\infty}\left((2\pi)\frac{v}{n}\exp\left(\frac{\left(\mu - \left(\frac{s_1}{n}\right)\right)^2}{\left(\frac{v}{n}\right)}\right)\right)^{(-1/2)} d\mu\left((2\pi)^{(n-1)}v^n n\exp\left(\frac{\left(\frac{s_2}{n}\right) - \left(\frac{s_1}{n}\right)^2}{\left(\frac{v}{n}\right)}\right)\right)^{(-1/2)}$$

and seeing that the remaining integral corresponds to a normal density with mean $\frac{s_1}{n}$ and variance $\frac{v}{n}$ it integrates to 1 and we obtain

$$p(n,s_1,s_2)=\left((2\pi)^{(n-1)}v^n n\exp\left(\frac{\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2}{\left(\frac{v}{n}\right)}\right)\right)^{(-1/2)}$$

as our posterior normalization constant as a function of the sufficient statistics s_1 and s_2.

Seeking the Bayesian predictive distribution we take

$$g(x|n,s_1,s_2)=\frac{p(n+1,s_1+x,s_2+x^2)}{p(n,s_1,s_2)}=\left((2\pi)v\left(\frac{n+1}{n}\right)\exp\left(\frac{\left(\frac{s_2+x^2}{n+1}\right)-\left(\frac{s_1+x}{n+1}\right)^2}{\left(\frac{v}{n+1}\right)}-\frac{\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2}{\left(\frac{v}{n}\right)}\right)\right)^{(-1/2)}$$

which while correct is not that useful so we to get

$$g(x|n,s_1,s_2)=\left((2\pi)v\left(\frac{n+1}{n}\right)\exp\left(\frac{x^2n^2-2xs_1n+s_1^2}{vn(n+1)}\right)\right)^{(-1/2)}$$

which we then simplify into

$$g(x|n,s_1,s_2)=\left((2\pi)v\left(\frac{n+1}{n}\right)\exp\left(\frac{\left(x-\frac{s_1}{n}\right)^2}{v\frac{(n+1)}{n}}\right)\right)^{(-1/2)}$$

or in terms of the maximum likelihood estimator $\hat{\mu}_{ML}$

$$g(x|n,s_1,s_2)=\left((2\pi)v\left(\frac{n+1}{n}\right)\exp\left(\frac{\left(x-\hat{\mu}_{ML}\right)^2}{v\frac{(n+1)}{n}}\right)\right)^{(-1/2)}$$

which corresponds to a normal distribution with a mean of $\hat{\mu}_{ML}$ and variance $v(n+1)/n$.

And with the Bayesian predictive density known we have everything we need to calculate DKL_{ML} and DKL_B for a specific sample where

$$DKL_{ML}=\int_{-\infty}^{\infty}f(x|\mu_0,v_0)\log\left(\frac{f(x|\mu_0,v_0)}{f(x|\hat{\mu}_{ML},v_0)}\right)dx=\frac{(\mu_0-\hat{\mu}_{ML})^2}{2v_0}$$

$$DKL_B=\int_{-\infty}^{\infty}f(x|\mu_0,v_0)\log\left(\frac{f(x|\mu_0,v_0)}{f(x|\hat{\mu}_{ML},v_0(n+1)/n)}\right)dx=\frac{(\mu_0-\hat{\mu}_{ML})^2}{2v_0(n+1)/n}+\frac{1}{2}\left(\frac{-1}{n+1}+\log\left(\frac{n+1}{n}\right)\right)$$

thanks to both predictive densities being normal. From this we note that both KL divergences are functions of $\hat{\mu}_{ML}$ whose marginal distribution is a normal with mean μ_0 and variance v_0/n. Using $f(\hat{\mu}_{ML})$ here as the marginal density of $\hat{\mu}_{ML}$ we obtain

$$EKL_{ML}=\int_{-\infty}^{\infty}f(\hat{\mu}_{ML}|\mu_0,v_0/n)\frac{\left(\hat{\mu}_{ML}-\mu_0\right)^2}{2v_0}d\hat{\mu}_{ML}=\frac{1}{2n}$$

$$EKL_B=\int_{-\infty}^{\infty}f(\hat{\mu}_{ML}|\mu_0,v_0/n)\frac{\left(\hat{\mu}_{ML}-\mu_0\right)^2n}{2v_0(n+1)}-\frac{1}{2(n+1)}+\frac{1}{2}\log\left(\frac{(n+1)}{n}\right)d\hat{\mu}_{ML}=\frac{1}{2}\log\left(\frac{(n+1)}{n}\right)$$

where we see that EKL_B is smaller then EKL_{ML} for all positive integers.

4.3. Unknown Mean and Unknown Variance

From the square root of the determinant of the Fisher information matrix we have Jeffreys's prior as $\pi_0(\mu, v) = (2v^3)^{(-1/2)}$ and the posterior normalization constant as

$$p(n, s_1, s_2) = \int_0^{\infty}\int_{-\infty}^{\infty} f(x_{1:n}|\mu, v)\pi_0(\mu, v)d\mu dv = \int_0^{\infty}\int_{-\infty}^{\infty}\left(\exp\left(\frac{\mu^2 n - 2\mu s_1 + s_2}{v} + \log(2(2\pi)^n v^{n+3})\right)\right)^{(-1/2)} d\mu dv$$

which we can split into

$$p(n, s_1, s_2) = \int_0^{\infty}\int_{-\infty}^{\infty}\left(\exp\left(\frac{n\left(\mu - \frac{s_1}{n}\right)^2}{v} + \log\left(2\pi\frac{v}{n}\right)\right)\exp\left(\frac{n\left(\left(\frac{s_2}{n}\right) - \left(\frac{s_1}{n}\right)^2\right)}{v} + \log\left(2(2\pi)^{n-1}v^{n+2}n\right)\right)\right)^{-1/2} d\mu dv$$

where the right hand part of the expression can be moved out of the inner integral leaving the left hand expression as the typical normal density integrating to 1. With the inner integral gone we now have

$$p(n, s_1, s_2) = \int_0^{\infty}\exp\left(\frac{n\left(\left(\frac{s_2}{n}\right) - \left(\frac{s_1}{n}\right)^2\right)}{2v} + \left(\frac{n}{2}+1\right)\log(v) + \left(\frac{1}{2}\right)\log\left(2(2\pi)^{n-1}n\right)\right)^{-1} dv$$

which with that substitutions $\alpha = n/2$ and $\beta = n((s_2/n) - (s_1/n)^{2)}/2$ can be rewritten into

$$p(n, s_1, s_2) = \int_0^{\infty}\exp\left(\frac{\beta}{v} + (\alpha+1)\log(v) - \alpha\log(\beta) + \log(\Gamma(\alpha))\right)^{-1}\exp\left(\alpha\log(\beta) - \log(\Gamma(\alpha)) + \log(2(2\pi)^{n-1}n)/2\right)^{-1} dv$$

where the left hand expression corresponds to the density of an inverse-gamma distribution and the right hand side expression does not depend on v and therefore moved out with the remaining expression within the integral integrating to 1. Simplifying the remaining expression gives us the posterior normalization constant as a function of the sufficient statistics s_1 and s_2

$$p(n, s_1, s_2) = \Gamma\left(\frac{n}{2}\right)\left(\left(\left(\frac{s_2}{n}\right) - \left(\frac{s_1}{n}\right)^2\right)^n \pi^{n-1} n^{n+1}\right)^{-(1/2)}$$

from which we calculate the Bayesian predictive density

$$g(x|n, s_1, s_2) = \frac{p(n+1, s_1+x, s_2+x^2)}{p(n, s_1, s_2)} = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{\left(\left(\frac{s_2+x^2}{n+1}\right) - \left(\frac{s_1+x}{n+1}\right)^2\right)^{(n+1)}}{\left(\left(\frac{s_2}{n}\right) - \left(\frac{s_1}{n}\right)^2\right)^{(n)}}\pi\frac{(n+1)^{(n+2)}}{n^{(n+1)}}\right)^{-(1/2)}$$

which while correct is terribly unintuitive so we simplify it into

$$g(x|n,s_1,s_2)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n+1}}\left(\left(\frac{\left(\frac{s_2+x^2}{n+1}\right)-\left(\frac{s_1+x}{n+1}\right)^2}{\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2}\right)^{(n+1)}\left(\frac{n+1}{n}\right)^{(n+1)}\pi\left(\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2\right)\right)^{-(1/2)}$$

which is still horrible so we simplify more to get

$$g(x|n,s_1,s_2)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n+1}}\left(\left(\frac{(s_2n-s_1^2)+x^2n-2xs_1+s_2}{s_2n-s_1^2}\right)^{(n+1)}\left(\frac{n}{n+1}\right)^{(n+1)}\pi\left(\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2\right)\right)^{-(1/2)}$$

which we will continue simplifying

$$g(x|n,s_1,s_2)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n+1}}\left(\left(\frac{n+1}{n}+\frac{(x-s_1/n)^2}{n\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2}\right)^{(n+1)}\left(\frac{n}{n+1}\right)^{(n+1)}\pi\left(\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2\right)\right)^{-(1/2)}$$

which we finally simplify into

$$g(x|n,s_1,s_2)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n+1}}\left(\left(1+\frac{\left(x-\frac{s_1}{n}\right)^2}{(n+1)\left(\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2\right)}\right)^{(n+1)}\pi\left(\left(\frac{s_2}{n}\right)-\left(\frac{s_1}{n}\right)^2\right)\right)^{(-1/2)}$$

Alternatively, we express the Bayesian predictive density in terms of the maximum likelihood estimators $\hat{\mu}_{ML}$ nad $\hat{v}_{ML}$

$$g(x|n,\hat{\mu}_{ML},\hat{v}_{ML})=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n+1}}\left(\left(1+\frac{(x-\hat{\mu}_{ML})^2}{(n+1)\hat{v}_{ML}}\right)^{(n+1)}\pi\hat{v}_{ML}\right)^{(-1/2)}$$

which is interestingly quite similar to the density of the student t distribution.

5. Conclusion

We conclude this thesis by summarizing the two main results; that for single parameter exponential models and unknown mean known variance normal models the Bayesian predictive distribution outperforms the ML predictive distribution in terms of expected KL divergence between the true and predictive densities, and the Bayesian predictive densities also outperform the asymptotic KL divergence result for all sample sizes. The technique of dividing posterior normalization constants was also used to derive the predictive density for an exponential distribution and for normal distributions, both unknown mean with known variance and unknown mean with unknown variance. It was also noted that the Bayesian predictive density function could in all these cases be conveniently parameterized by the MLE instead of the sufficient statistics.

Example of potential future work would be to determine the Bayesian predictive distributions for more models, determine if the observed superiority over ML predictive distributions holds more generally, and if Jeffreys's prior is the optimal equivariant prior if one seeks to minimize average KL divergence. Calculating higher order cumulants, like the variance, of the KL divergence would also be of interest.

As a final statement I will conjecture that the Bayesian predictive distribution when used with Jeffreys's prior always outperforms the ML predictive distribution in terms of minimizing the expected KL divergence and that Jeffreys's prior is the best equivariant prior for minimizing it.

References

Abramowitz & Stegun (1964)
Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables
Edited by: Milton Abramowitz and Irene A. Stegun
Publisher: United States Department of Commerce, National Bureau of Standards (NBS)
URL: https://www.archive.org/details/AandS-mono600
Notes: The NBS is today known as National Institute of Standards and Technology (NIST).

Alzer (1997)
On some inequalities for the Gamma and Psi functions
Author: Horst Alzer
Journal: Mathematics of Computation (volume 66, pages 373-389)
URL: https://www.ams.org/journals/mcom/1997-66-217/S0025-5718-97-00807-7/S0025-5718-97-00807-7.pdf

Appendix A - Complimentary Proofs

Ola's Asymptotic Distribution Independent Result

The proof of Ola's asymptotic result is simple to obtain using the second order Taylor expansion on the definitions of EKL_{ML} nad EKL_B, which by using $f(\hat{\lambda}_{ML})$ as the marginal density of $\hat{\lambda}_{ML}$ becomes

$$\int_{\hat{\lambda}_{ML-}}^{\hat{\lambda}_{ML+}} f(\hat{\lambda}_{ML})(\hat{\lambda}_{ML}-\lambda_0)^T \frac{I(\hat{\lambda}_0)}{2}(\hat{\lambda}_{ML}-\lambda_0)d\hat{\lambda}_{ML}$$

which together with the asymptotic normality with mean λ_0 and variance $(nI(\lambda_0))^{(-1)}$ becomes

$$(2n)^{(-1)}\sum_{i,j}(I(\hat{\lambda}_0)I(\hat{\lambda}_0)_{i,j}^{(-1)})$$

where the Fisher information matrices cancels each other out and the expression simplifies into $\frac{p}{2n}$, where $p = dim(\hat{\lambda}_{ML})$, the asymptotic result.

The second order Taylor expansion is however not a well know result for the Bayesian predictive distribution so for completeness we take the definition of DKL_B and the Bayesian predictive density

$$DKL_B(n) = \int_{\hat{\lambda}_{ML-}}^{\hat{\lambda}_{ML+}} f(x|\hat{\lambda}_0)\log\left(\frac{f(x|\hat{\lambda}_0)}{\int_{\lambda_-}^{\lambda_+} f(x|\lambda)\pi(\lambda|\hat{\lambda}_{ML})d\lambda}\right)d\hat{\lambda}_{ML}$$

Now we will do three thing: Replace $f(x|\lambda)$ with its second order Taylor expansion around λ_0, utilise the Bernstein-von Mises theorem to solve the inner integral using a normally distributed posterior, and simplify the above into

$$DKL_B(n) = \int_{x_-}^{x_+} f(x|\hat{\lambda}_0)\left(-\frac{f'_\lambda(x|\lambda_0)}{f(x|\lambda_0)}(\hat{\lambda}_{ML}-\lambda_0)+\left(\frac{f'_\lambda(x|\lambda_0)}{f(x|\lambda_0}\right)^2\frac{(\hat{\lambda}_{ML}-\lambda_0)^2}{2}-\frac{f'{}'_\lambda(x|\lambda_0)}{f(x|\lambda_0)}\frac{(\hat{\lambda}_{ML}-\lambda_0)^2}{2}+o((\hat{\lambda}_{ML}-\lambda_0)^2)\right)d\hat{\lambda}_x$$

which can be solved for

$$DKL_B(n) = (\hat{\lambda}_{ML}-\lambda_0)^T\frac{I(\hat{\lambda}_0)}{2}(\hat{\lambda}_{ML}-\lambda_0)+o((\hat{\lambda}_{ML}-\hat{\lambda}_0)^2)$$

which lets us apply the above proof for the asymptotic result for EKL_{ML} for EKL_B as well.

First Predictive Distribution KL Divergence Integral

We seek to prove that

$$\exp(t_1)\Gamma(0,t_1) = \int_0^\infty \exp(-z)\log\left(1+\frac{z}{t_1}\right)dz$$

so we take the right hand integral and perform integration by parts to get

$$\int_0^\infty \exp(-z)\log\left(1+\frac{z}{t_1}\right)dz = \left[-\exp(-z)\log\left(1+\frac{z}{t_1}\right)\right]_0^\infty - \int_0^\infty \exp(-z)\left(\frac{1}{t_1+z}\right)dz$$

with the middle part having trivial limits of 0 in both directions, so that the whole expression simplifies to

$$\int_0^{\infty} -\exp(-z)\left(\frac{1}{t_1+z}\right)dz$$

which can be rewritten to

$$\exp(t_1)\int_0^{\infty} -\exp(-(t_1+z))\left(\frac{1}{t_1+z}\right)dz$$

and by the substitution $y = t_1 + z$ results in

$$\exp(t_1)\int_{t_1}^{\infty} \exp\left(-y\right)\left(\frac{1}{y}\right)dy$$

Through the definition of the upper incomplete gamma function the last displayed equation becomes

$$\exp(t_1)\Gamma(0, t_1)$$

which was to be proven, see Abramowitz & Stegun (1964) page 262 (6.5.15).

Gamma Logarithmic Moment

We seek to prove that assuming a gamma random variable Z with density $f(z|\alpha, \beta)$, we have

$$\int_0^{\infty} f(z|\alpha, \beta)\log(z)dz = \psi(\alpha) - \log(\beta),$$

where $\psi(x)$ is the digamma function defined as $\psi(x) = \Gamma'(x)/\Gamma(x)$, see Abramowitz & Stegun (1964) page 258 (6.3.1).

We take the left hand side and expand the density knowing its integral is 1

$$1 = \int_0^{\infty} \exp\left(-\beta z + (\alpha-1)\log(\beta z) + \log(\beta)\right)/\Gamma(\alpha)dz$$

Then we multiply both sides by $\Gamma(\alpha)$ and take the derivative with regard to α

$$\Gamma'(\alpha) = \int_0^{\infty} \log(\beta z)\exp\left(-\beta z + (\alpha-1)\log(\beta z) + \log(\beta)\right)dz$$

Then we divide both sides by $\Gamma(\alpha)$ and substitute back the density

$$\Gamma'(\alpha)/\Gamma(\alpha) = \int_0^{\infty} f(z|\alpha, \beta)\log(\beta z)dz$$

in which we use the definition of the digamma function and splitting the logs to finaly get

$$\psi(\alpha) - \log(\beta) = \int_0^{\infty} f(z|\alpha, \beta)\log(z)dz$$

which was to be proven.

Gamma Inverse Moment

We seek to prove that assuming a gamma random variable Z with density function $f(z|\alpha, \beta)$, we have

$$\int_0^{\infty} f(z|\alpha, \beta)z^{-1}dz = \beta\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}$$

We start with the left hand side integral and expand the density function into

$$\int_0^\infty z^{-1} \exp(-\beta z) z^{\alpha-1} \beta^\alpha / \Gamma(\alpha) dz$$

and rewrite it into

$$\beta \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)} \int_0^\infty \exp(-\beta z) z^{\alpha-2} \beta^{\alpha-1} / \Gamma(\alpha-1) dz$$

witch with the substitution $\alpha' = \alpha - 1$ can be rewritten into

$$\beta \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)} \int_0^\infty f(z|\alpha', \beta) dz$$

with the density integrating to 1 resulting in

$$\beta \frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}$$

which was to be proven.

First Predictive Distribution KL Divergence Integral

We seek to prove that

$$\int_0^\infty f(t_1) \left(\log(t_1) - \log(n) - 1 + (n+1) \exp(t_1) \Gamma(0, t_1) \right) dt_1 = \psi(n) - \log(n) + \frac{1}{n}$$

which we will do by splitting the integral into

$$\int_0^\infty f(t_1) \log(t_1) dt_1 - (\log(n) + 1) \int_0^\infty f(t_1) dt_1 + (n+1) \int_0^\infty f(t_1) \exp(t_1) \Gamma(0, t_1) dt_1 = \psi(n) - \log(n) + \frac{1}{n}$$

and substituting the logarithmic moment into $\psi(n) - \log(n)$ leaving

$$(n+1) \int_0^\infty f(t_1) \exp(t_1) \Gamma(0, t_1) dt_1 + \psi(n) - \log(n) - 1 = \psi(n) - \log(n) + \frac{1}{n}$$

which we simplify into

$$\int_0^\infty f(t_1) \exp(t_1) \Gamma(0, t_1) dt_1 = \frac{1}{n}$$

leaving only the above moment of t_1 left to prove.

To prove the last displayed formula, we begin by noting that

$$\frac{d}{dt_1} \Gamma(0, t_1) = \frac{exp(-t_1)}{-t_1}$$

which is trivial to obtain using

$$\frac{d}{dx} \Gamma(a, x) = -\frac{\exp(-x)}{x^{a-1}}$$

from Abramowitz & Stegun (1964) page 262 (6.5.25) with $a = 0$.

Going back to the moment to prove, expand the density and simplify the integral into

$$\int_0^\infty t_1^{n-1}\Gamma(0,t_1)/\Gamma(n)dt_1$$

and performing integration by parts to get

$$\frac{[t_1^n\Gamma(0,t_1)]_0^\infty}{n\Gamma(n)} + \frac{1}{n}\int_0^\infty \frac{t_1^{n-1}\exp(-t_1)}{\Gamma(n)}\,dt_1$$

and note that the integral corresponds to $f(t_1)$ which integrates to 1 to get

$$\frac{[t_1^n\Gamma(0,t_1)]_0^\infty}{n\Gamma(n)} + \frac{1}{n}$$

leaving only the left side of the expression to prove 0.

We now as the final step seek to prove $[t_1^n\Gamma(0,t_1)]_0^\infty = 0$ which we will do by proving each limit to individually be zero.

Firstly from using the definition of $\Gamma(0,t_1)$ we have for $0 < x < 1$

$$|\Gamma(0,x)| \le |\Gamma(0,1)| + \int_x^1 t^{-1}dt = |\Gamma(0,1)| + \log(|x^{-1}|)$$

from which we get

$$x^n|\Gamma(0,x)| \le x^n|\Gamma(0,1)| + x^n\log(|x^{-1}|)$$

where the right hand side goes to 0 as $x \to 0$ for all positive integer n. Knowing that sample sizes are all positive integers we use to above to see that $t_1^n|\Gamma(0,t_1)| \to 0$ as $t_1 \to 0$.

For the second limit $t_1 \to \infty$ we use again the definition of $\Gamma(0,t_1)$ to get the inequality

$$|\Gamma(0,x)| \le x^{-1}\int_x^\infty \exp(-t)dt = x^{-1}\exp(-x)$$

which gives us

$$x^n|\Gamma(0,x)| \le x^{n-1}\exp(-x)$$

where the right hand side goes to 0 as $x \to \infty$ which gives us a stronger result then the sought $t_1\Gamma(0,t_1) \to 0$ as $t_1 \to \infty$.

From applying the above two results we have $[t_1\Gamma(0,t_1)]_0^\infty = 0$ which mean we finally have our sought result of

$$\int_0^\infty f(t_1)\exp(t_1)\Gamma(0,t_1)dt_1 = \frac{1}{n}$$

for proving that $EKL_B = \psi(n) - \log(n) + 1/n$