Algorithmic Inference of Two-Parameter Gamma
Distribution

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Communications in Statistics - Simulation and Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>LSSP-2009-0057.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>01-Jul-2009</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Apolloni, Bruno; University of Milano, Department of Computer Science</td>
</tr>
<tr>
<td></td>
<td>Bassis, Simone; University of Milano, Department of Computer Science</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Two-parameters Gamma distribution, Algorithmic inference, Population bootstrap, Gibbs sampling, Adaptive rejection sampling, Maximum likelihood estimator</td>
</tr>
<tr>
<td>Abstract:</td>
<td>We provide an estimation procedure of the two-parameter Gamma distribution based on the Algorithmic Inference approach. As a key feature of this approach, we compute the joint probability distribution of these parameters without assuming any prior. To this end we propose a numerical algorithm which is often beneficial of a highly efficient speed up based on an approximate analytical expression of the probability distribution. We contrast our interval and point estimates with those recently obtained in Son and Oh (2006). We realize that our estimates are both unbiased and more accurate, albeit more dispersed than Bayesian methods.</td>
</tr>
</tbody>
</table>

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

URL: http://mc.manuscriptcentral.com/lssp E-mail: comstat@univmail.cis.mcmaster.ca
infgamma.zip
INFERENVE

ALGORITHMIC INFERENC   OF TWO-PARAMETER GAMMA DISTRIBUTION

B. Apolloni and S. Bassis
Dipartimento di Scienze dell’Informazione
University of Milano
Via Comelico 39/41, 20135
Milano, Italy
{apolloni, bassis}@dsi.unimi.it

Key Words: Two-parameters Gamma distribution; Algorithmic inference; Population bootstrap; Gibbs sampling; Adaptive rejection sampling; Maximum likelihood estimator.

ABSTRACT

We provide an estimation procedure of the two-parameter Gamma distribution based on the Algorithmic Inference approach. As a key feature of this approach, we compute the joint probability distribution of these parameters without assuming any prior. To this end we propose a numerical algorithm which is often beneficial of a highly efficient speed up based on an approximate analytical expression of the probability distribution. We contrast our interval and point estimates with those recently obtained in Son and Oh (2006) for the same problem. From this benchmark we realize that our estimates are both unbiased and more accurate, albeit more dispersed, in some cases, than the competitor methods’, where the dispersion drawback is notably mitigated w.r.t. Bayesian methods by a greater estimate decorrelation. We also briefly discuss the theoretical novelty of the adopted inference paradigm which actually represents a brush up on a Fisher perspective dating to almost a century, made feasible today by the available computational tools.

1. INTRODUCTION

The Gamma distribution is a formidable touchstone for any inference framework. On
the one hand it is widely employed in many application fields, ranging from reliability (Juraj and Godfrey, 1999; Corp., 2009) to telecommunications (Titus and Wheatley, 1996), insurance (Ammeter, 1970) and so on. On the other, the estimation of its parameters still poses a challenging problem. Adopting an analogous notation as in Son and Oh (2006), we introduce the Gamma distribution through its probability density function (PDF) applied to the random variable $X$ as follows:

$$f_X(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad x > 0, \quad \alpha, \beta > 0,$$

where $\alpha$ and $\beta$ are respectively a shape and scale parameter. Of the conventional point estimators, the one based on the method of moments (MM) is very simple but inefficient as well — with some exceptions in asymptotic conditions (Wiens et al., 2003). Maximum likelihood (ML) estimators may rely on a pair of joint sufficient statistics, but they are penalized by the absence of a closed analytical expression. To overcome this drawback, we find in the literature either approximate expressions (Thom, 1958; Greenwood and D., 1960) or recursive procedures to get a numerical solution (Bowman and Shenton, 1988). With the latter we work with a sequence of parameter values converging with some dynamics to the planned estimators. With a dual perspective, other researchers adopt the Bayes approach to collect a population of parameter estimators — possibly generated through a Gibbs sampler (Ripley, 1988) or more sophisticated techniques (Gilks and P., 1992) — and deduce point estimators as central values of this population.

Our work goes in the same direction, but without requiring to set any $a$ priori distribution of the parameters. We have implemented this kind of procedure in many inference problems, mainly within the machine learning framework (Apolloni et al., 2008a) and in specific operational fields such as survival analysis (Apolloni et al., 2007b), human mobility modeling (Apolloni et al., 2007a) and statistical tests of hypotheses (Apolloni and Bassis, 2007). The goal of this paper is twofold: on the one hand we want to efficiently solve the

---

1By default, capital letters (such as $U$, $X$) will denote random variables and small letters ($u$, $x$) their corresponding realizations; vectorial versions of the above variables will be denoted through bold symbols ($\mathbf{X}$, $\mathbf{U}$, $\mathbf{x}$, $\mathbf{u}$).
questioned estimation problem as a distinguished inference task. To this aim, we show our procedures to quickly compute accurate and unbiased estimators based on independent approximations of the unknown parameters. On the other hand, we would use this solution as a business card to introduce our inference framework, which we call Algorithmic Inference, to a more probability methodologically oriented audience and invite the reader to read more widely about it in a couple of books we have published on the subject (Apolloni et al., 2008c, 2006). Therefore the paper is organized as follows. In Section 2 we introduce our method from a strictly operational perspective, through definitions and a few logical links, having the questioned inference problem as lead example. In Section 3 we apply the proposed procedure to the same benchmark as Son and Oh (2006), contrasting our results with the competitor methods considered therein. We devote Section 4 to a discussion of the benefits of our approach and its connections to some inference strategies animating the debate since the early 20th century, to conclude the paper.

2. COMPATIBLE PARAMETERS

Given a distribution law, think of its unknown parameters as random parameters. But, in place of a specific prior distribution, they inherit probabilities from a standard source of random seeds that are also at the basis of the random variables they are deputed to characterize. Like with a barrel with two interlocking taps, through the former you get, from the source, samples of the random variable for given parameters; while from the latter you get samples of the parameters for given observed variables. You measure what happens with the former and compute what would happen with the latter. The basic steps of this twisting task are the following:

1. **Sampling mechanism** $\mathcal{M}_X$. It consists of a pair $(Z, g_\theta)$, where the seed $Z$ is a random variable without unknown parameters, while the *explaining function* $g_\theta$ is a function mapping from samples $\{z_1, \ldots, z_m\}$ of $Z$ to samples $\{x_1, \ldots, x_m\}$ of the random variable $X$ we are interested in. This function is indexed in $\theta$ which represents the not yet set parameters of the random variable, i.e. either a scalar value or a vector of values we want to investigate. Thanks to the probability integral transformation theorem (Rosenblatt,
For Peer Review Only

1952) we have that, by using the uniform variable $U$ in $[0, 1]$ as a seed, the explaining function $g_\theta$ for $X$ distributed according to (1) computes the following mapping:

$$x_i = F_{X}^{-1}(u_i)$$  (2)

where $F_{X}^{-1}$ is the inverse $h_{\alpha,\beta}$ of the Gamma cumulative distribution function (CDF). What is interesting is that $\beta$ plays the role of a scale factor, such that $h_{\alpha,\beta} = \beta h'_{\alpha}$, with $h'_{\alpha} = h_{\alpha,1}$. In turn $h'_{\alpha}$ has an analytical, though complex, expression that may be managed by widespread tools of symbolic computation. For instance in Mathematica (Wolfram Research Inc., 2008) it is computed by the routine InverseGammaRegularized[$\alpha, 0, u$] where $u$ is the argument of $F_{X}^{-1}$. Not always $U$ is the most appropriate seed. For instance one prefers using the standard Gaussian variable $\Psi$ as a seed of a Gaussian variable $X$ with mean $\mu$ and standard deviation $\sigma$ through the explaining function $x_i = (\mu + \psi_i \sigma)$, as $F_{X}^{-1}$ is unavailable in its exact form for this $X$.

All this falls in the common practice of random variable simulation. The benefit of formalizing it in terms of sampling mechanism $\langle Z, g_\theta \rangle$ lies firstly in a clear partition of what is out of our handling and what may be the object of our inference. We can say nothing new about seeds $\{z_1, \ldots, z_m\}$, since they are randomly tossed from a perfectly known distribution; hence they are completely unquestionable as for the single value, and completely known as for their ensemble properties. On the contrary, the explaining function groups into $\theta$ the free parameters that we want to infer from the sample. As for a second benefit, they are exactly the seeds that state links between observations and parameters of a given $X$. We cannot say which value has $\theta$, since we do not know the seeds of the observations. Rather, we may transfer the probability mass of the seeds from the sample to the parameters realizing the sample, as we will see in the next point.

2. Master equations. The actual connection between the model and the observed data is tossed in terms of a set of relations between statistics on the data and unknown parameters that come as a corollary of the sampling mechanism. We call these relations master
equations. Pivoting around the statistic \( s = h(x_1, \ldots, x_m) = h(g_\theta(z_1), \ldots, g_\theta(z_m)) \), where \( s \) and \( h \) are vectors in their general instantiation, the general form of a system of master equations is:

\[
s = \rho(\theta; z_1, \ldots, z_m).
\]  

(3)

With this relation we may inspect the values of the parameter \( \theta \) that could have generated a sample with the observed statistic \( s \) from a particular setting of the seeds \( \{z_1, \ldots, z_m\} \). Hence, if we draw seeds according to their known distribution – uniform in our case – we get a sample of parameters in response (Apolloni et al., 2006). In order to ensure this sample clean properties, we involve sufficient statistics w.r.t. the parameters (Stigler, 1973) in the master equations. In greater detail, with Gamma distributions we refer to the jointly sufficient statistics \( s_1 = \sum_{i=1}^m x_i \) and \( s_2 = \prod_{i=1}^m x_i \), which we prefer to represent through \( s_1 \) and \( s_3 = s_2/s_1^m \) so that the master equation on \( s_3 \) depends only on \( \alpha \). Namely, we will use the following set of master equations:

\[
s_1 = \beta \sum_{i=1}^m h'_\alpha(u_i)
\]

(4)

\[
s_3 = \frac{\prod_{i=1}^m h'_\alpha(u_i)}{(\sum_{i=1}^m h'_\alpha(u_i))^m}
\]

(5)

3. Parameters population. Having fixed a set of master equations, you may draw seeds in an infinitely large number so as to map from a population of seeds into a population that is representative of the random parameter \( \Theta \). The specific features of the mapping translate the uniform distribution of the former into a properly shaped distribution of the latter. In this way we obtain the graph in Figure 1 which reports the joint empirical distribution law of the random parameters \( A \) and \( B \) on the basis of the male mice dataset of Gross and Clark (1975) (G&C dataset henceforth), consisting of the 20 randomly selected survival times of male mice exposed to 240 rads of gamma radiation as listed in Table 1.

On the one hand we have no problem realizing that the mentioned probability distribution transform may occur, provided that some properties are owned by the involved statistics,
Table 1: Male mice dataset of Gross and Clark (1975).

<table>
<thead>
<tr>
<th>152</th>
<th>152</th>
<th>115</th>
<th>109</th>
<th>137</th>
<th>88</th>
<th>94</th>
<th>77</th>
<th>160</th>
<th>165</th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>40</td>
<td>128</td>
<td>123</td>
<td>136</td>
<td>101</td>
<td>62</td>
<td>153</td>
<td>83</td>
<td>69</td>
</tr>
</tbody>
</table>

Figure 1: Empirical joint density function of the $A$ and $B$ parameters of Gamma distribution laws that are compatible with G&C dataset, on the basis of 10,000 bootstrap replicas.

such as the univocal invertibility of (3) w.r.t. $\theta$ or the feasibility of the inversion solution. We formally denoted a statistic with these properties as well behaving in Apolloni et al. (2008c) and realized that a statistic whose components are jointly sufficient is well behaving. On the other hand, the value of the new distribution stands in measuring a compatibility feature of the found parameters with the observed statistic that we formalize through the following definition.

**Definition 1.** For a random variable $X$ and a sample drawn from it, a compatible distribution is a distribution having the same sampling mechanism $M_X = (Z, g_{\hat{\theta}})$ of $X$ with a value $\hat{\theta}$ of the parameter $\Theta$ derived from a system of master equations (3) rooted on a well behaving statistic $s$. In turn, $\hat{\theta}$ is a $\Theta$ value compatible with the sample.

As for the algorithms implementing the above procedure, we distinguish a numerical implementation from an analytical counterpart, leaving room for any intermediate specification. The lead example developed till now is an instance of the following Algorithm 1 that we call $p$-bootstrap in order to both evoking the original idea of Efron and Tibshirani (1993)
and highlighting the fact that we bootstrap population parameters and not samples.

**Algorithm 1** Generating parameter populations through bootstrap.

1. Identify a well behaving statistic $S$ for the parameter $\theta$ and the related system of master equations;

2. compute a realization $s_\Theta$ of $S$ from a sample of size $m$;

3. repeat for a satisfactory number $n$ of iterations:
   
   (a) draw a sample $\tilde{z}_i$ of size $m$ from the seed deputed random variable;
   
   (b) get $\tilde{\theta}_i = \text{Inv}(s_\Theta, \tilde{z}_i)$ as a solution of (3) in $\tilde{\theta}$ with $s = s_\Theta$ and $(z_1, \ldots, z_m) = \tilde{z}_i$;
   
   (c) add $\tilde{\theta}_i$ to $\Theta$ population.

Since we have no limits to the number of seed replicas – apart from the computational time we are allowed to spend – we may approximate with any accuracy probabilities of events related to the parameters through the companion frequencies. We just highlight the different role of the sample size decreeing the shallowness of the parameter distribution, from the number of replicas determining the accuracy with which this distribution is computed.

Figure 1 was obtained exactly with this procedure, where we numerically realize that inversion of (5) is univocal since $s_3$ results to be a monotonically increasing function of $\alpha$ for whatever seeds $\{u_1, \ldots, u_m\}$ we checked.

Things go better, as for computational loads, if we may identify the analytical form of the above distribution. In our case, the two master equations together with the well behaving-ness of $s_1$ and $s_3$ guarantee that for both distributions we may state the following relationships:

\[ s_1 \leq \tilde{s}_1 \iff \beta \leq \tilde{\beta} \quad \forall \alpha \quad (6) \]

\[ s_3 \leq \tilde{s}_3 \iff \alpha \leq \tilde{\alpha} \quad (7) \]

where $\tilde{s}_i$ is the value of the statistic we would have observed if the parameter had moved to $\tilde{\beta}$ in the master equation (3) for the same specifications $\{z_1, \ldots, z_m\}$ of $Z$ and the same $\alpha$, and
similarly for $\tilde{s}_3$ but independently of $\beta$. Summing up these implications over all seeds and maintaining fixed $s_1, s_3, \tilde{\alpha}, \tilde{\beta}$, we have analogous relationships between events (see Apolloni et al. (2006) for a wider discussion):

$$ (s_1 \leq \tilde{S}_1) = (B \leq \tilde{\beta}) \quad \forall \alpha $$

$$ (s_3 \leq \tilde{S}_3) = (A \leq \tilde{\alpha}) $$

Hence, moving to probability we have:

$$ F_{B|A=\alpha}(\beta) = 1 - F_{S_1}(s_1) = Q(\alpha m, s_1/\beta) $$

$$ F_A(\alpha) = 1 - F_{S_3}(s_3) \sim Q(\phi_1(\alpha), s_3 \phi_2(\alpha)) $$

where $Q(a, z)$ denotes the regularized incomplete Gamma function (Chaudhry and Zubair, 2001) and $\phi_1, \phi_2$ are suitable functions we are going to introduce. Namely, while $S_1$ follows a Gamma distribution with parameters $m\alpha$ and $\beta$, the distribution of $S_3$ has no immediate analytical closed form. If we read it in terms of the product of Dirichlet components, its distribution – while treatable for $m = 2$ (Nadarajah and Kotz, 2004) – becomes computationally costly with higher sample sizes, mainly due to its characterization in terms of the Fox’s H-function (Garg et al., 2002) or its derivatives (Coelho and Mexia, 2007). Thus, abandoning the search for an exact analytical form, we find again in the Gamma distribution a good candidate function for approximating $F_{S_3}$. Then, with the mentioned limits of MM estimators for this distribution, we exploit the relative ease of computing the first and second $S_3$ moments (Nadarajah and Kotz, 2004), and obtain an approximate distribution of $A$ by plugging in the above approximating function the MM estimators, say $(t_1, t_2)$, of its parameters:

$$ t_1 = \phi_1(\alpha) = \left( \frac{\Gamma(\alpha + 1)^2 \Gamma(\alpha + 2)^{2m} \Gamma(\alpha + m + 1)^2}{\Gamma(\alpha m) \Gamma((\alpha + 2)m - 2)} - 1 \right)^{-1} $$

$$ t_2 = \phi_2(\alpha) = \left( \frac{\Gamma(\alpha + 1)^{1-m} \Gamma(\alpha + 2)^{m-1} \Gamma(\alpha + m - 1)}{\Gamma((\alpha + 2)m - 2)} - \frac{\Gamma(\alpha)^{1-m} \Gamma(\alpha + 1)^{m-1} \Gamma(\alpha m)}{\Gamma(\alpha m + m - 1)} \right)^{-1} $$

Does it make sense to use MM to approximate the $A$ distribution law in order to overcome the drawbacks of the $A$ MM estimator? In Figure 2 we show that this $F_{S_3}$ approximation
may be very tight. On the one hand the shape of the Gamma CDF is well suited to the approximating purpose for $\alpha \geq 2$. On the other hand, however, we suffer from the bias of the MM estimators; therefore we must introduce some multiplicative coefficients $\gamma_1$ and $\gamma_2$ as tuning terms in (12) and (13). Since the ultimate goal is the approximation of $A$ distribution, we get them directly from the interpolation of the $A$ distribution function. Namely, denoting with $F_{A}^{-1}(r)$ the $r$-th quantile of $A$ distribution computed via $\mu$-bootstrap method and with $Q_{(a,z)}^{-1}(r)$ the analogous quantile using (11), we use the 0.25 and 0.75 quantiles to obtain the tuning coefficients $\tilde{\gamma}_i$s, i.e.:

$$
Q_{(\tilde{\gamma}_1, \tilde{\gamma}_2)}^{-1}(0.25) = F_{A}^{-1}(0.25) \quad (14)
$$

$$
Q_{(\tilde{\gamma}_1, \tilde{\gamma}_2)}^{-1}(0.75) = F_{A}^{-1}(0.75) \quad (15)
$$

We compute coefficients on a thousand of samples of a Gamma distribution with different sample size and distribution parameters. By interpolating these data, we obtain the general expression of $\gamma_1$ as a quadratic function of $\log s_3$ and $m$ which works well for all our purposes in this paper:

$$
\log \gamma_1 = 0.6476 - 0.1235m - 0.0077m^2 - 0.0548 \log s_3 - 0.0021m \log s_3 - 0.0001 \log^2 s_3 \quad (16)
$$
Since we did not achieve a similar success with $\gamma_2$, we decided to compute this coefficient directly in the single inference instances by fitting the empirical CDF obtained from a small set of $\alpha$ p-bootstrap replicas. With this strategy, the method we propose is a hybrid numerical-analytical procedure where the injection of the analytical knowledge, in terms of expressions (11) to (13), constitutes a speed up of the pure numerical solution allowing a notably reduction of the number of p-bootstrap replicas. Within the benchmarks used in this paper the narrowing rate ranges from 25 to 250. This entails an almost identical reduction in running times – since the time required to compute the above expression is negligible in comparison to the replicas’ generation time. Thus the hybrid algorithm represents a relatively quick procedure attaining a good compromise between fastness and accuracy of the estimation algorithms, as will be shown in the next section. Focusing on the single components of the parameter vector and related scalar statistic, we may standardize the analytical computation with the following Algorithm 2. The discretization granule $\Delta > 0$ therein accounts for

**Algorithm 2** Generating parameter populations through analytical arguments.

1. Identify a well behaving statistic $S$ for the parameter $\theta$ and its discretization granule $\Delta$ (if any);

2. decide the monotony versus;

3. compute $F_{\Theta | S=s}(\theta) \in (q_1(F_{S|\Theta=\theta}(s)), q_2(F_{S|\Theta=\theta}(s)))$ where:

\[
q_i(F_S) = \begin{cases} 
q_1 = q_2 & \text{if } S \text{ continuous} \\
q_1(F_S(s)) = q_2(F_S(s - \Delta)) & \text{if } S \text{ discrete}
\end{cases}
\]

\[
q_i(F_S) = \begin{cases} 
1 - F_S & \text{if } s \text{ does not decrease with } \theta \\
F_S & \text{if } s \text{ does not increase with } \theta
\end{cases}
\]

for $i = 1, 2$.

the $S$ value jumps in the case of discrete random variables. In this case, the confluence of
multiple seeds to the same sampling mechanism outputs induces an indeterminacy in the \( \Theta \) distribution law reckoned by the pair \((q_1, q_2)\) (see Apolloni et al. (2006) for wider discussion on the algorithm).

3. NUMERICAL EXPERIMENTS

Like dwarfs on the shoulders of giants, we decide to use the same benchmarks as in Son and Oh (2006). In principle we add a pair of new rows on the tables therein to contrast and compare our results with those of competitors, having the adaptive method proposed by the authors of the referred paper as the favorite benchmark.

Table 2 represents a set of reconstruction experiments. Starting with a set of Gamma variables with the same \( \beta = 1 \) and \( \alpha \) spanning in the set \( \{0.5, 1, 2, 3, 4, 5\} \), we generate a sample of 30 elements of each variable, on the basis of which we estimate the above parameters. Our competitors’ methods are: i) two moment method implementations, one in the basic version (MM), and the other in the variant proposed in Wiens et al. (2003) and with a special selection of free parameters (WCB); ii) three maximum likelihood implementations, respectively with the approximations proposed in Thom (1958) (ML\(_{\text{Thom}}\)), in Greenwood and D. (1960) (ML\(_{\text{GD}}\)), and with the recursive procedure introduced in Bowman and Shenton (1988) (ML\(_{\text{BS}}\)); and iii) two Bayesian methods based on the prior:

\[
f_{A,B}(\alpha, \beta) \propto 1/\beta
\]

and Gibbs sampling of parameters \( \alpha \) and \( \beta \) from the full conditional distributions. Namely, to achieve samples, a rejection method (Ripley, 1988) is used having a proposal distribution either normal (RS) (Tsionas, 2001) or adaptive (ARS) (Gilks and P., 1992). Bayesian methods generate samples of 1,000 realizations of the random parameters’ posterior distribution – after a burn-in period of 100 iterations – and assume their mean as an estimate of the original parameters. The statistical behavior of the estimators is assessed over 1,000 samples of the Gamma variable for each parameter setting. The mean and the standard deviation are computed over the corresponding parameter estimates. Son and Oh focus only on \( \alpha \) estimates, having the \( \beta \) estimation as a minor task. Indeed, with MM and ML methods
Table 2: First and second order statistics for a set of reconstruction experiments for different values of $\alpha$ and $\beta = 1$. Central value and Dispersion columns refer respectively to sample mean and standard deviation, with the exception of the AI approach, where median and median deviation are used.
this estimation descends straightforwardly from the following relation between \( \hat{\alpha} \) and \( \hat{\beta} \), as estimators of the corresponding parameters:

\[
\hat{\beta} = s_1 / (m \hat{\alpha})
\]

(18)

As to the Bayesian methods, for \( \alpha \) given, sampling \( \beta \) from its conditional posterior distribution looks trivial. The whole experimental campaign was repeated with samples of size 100. On our part, since we are working with interpreted languages, namely Mathematica (Wolfram Research Inc., 2008), we preferred to avoid bearing the heavy computational effort coming from the second set of simulations, in favor of specific experiments aimed at identifying the different features of our approach w.r.t. the other ones. Hence, in the mentioned table and with reference to the sole sample size 30, we add rows concerning central value and dispersion of estimators from our approach (AI) using both p-bootstrap method and its hybrid version.

On each sample we compute 1,000 bootstrap replicas \((\hat{\alpha}, \hat{\beta})\) of the parameters, as a counterpart of analogous values obtained with the Bayesian methods. The hybrid version uses subsamples of size 40 (just picking the first 40 of the above replicas) to compute \( \gamma_2 \). Unlike the Bayesian methods, we opt for the sample median in place of sample mean to identify a central value. Analogously, we use the median of the medians coming from the 1,000 samples as indicator of their central value. Indeed, we assume this statistic to be more suitable within the Algorithmic Inference approach. The problem is that, on the one hand with this approach we refer to the parameter distribution laws compatible with the observed sample. On the other, we generate a set of samples with exactly the same parameter \( \alpha \). Hence, to adapt our method to the reconstruction problem we must use a statistic that is substantially independent of the (either compatible or actual) parameter distribution law. This occurs with the median, since, in the mechanism generating parameter replicas from a sample \( \{x_1, \ldots, x_m\} \), the median of \( \hat{\alpha} \) depends only on the median of \( \sum_{i=1}^{m} \log(1 - u_i) \) and on no other detail of the distribution of this parameter. The same consideration holds for the median of the medians coming from the samples. Thus, this statistic is not biased by the actual parameter distribution law used to generate the sample, which is the
degenerate distribution with the support coinciding with one of the constant values of $\alpha$. On the contrary, the adopted constant value coincides with the median of the degenerate distribution, in turn converging in probability with the median of medians (Zhu and Wang, 2005), which we use therefore as a resuming central value statistic.

As a result, we obtain in Table 2 central values that in 5 cases out of 6 with p-bootstrap and 3 cases out of 4 with the hybrid version are closer to the original parameter than those provided by ARS, and even better w.r.t. other competitors’ methods. Above all, our estimates are unbiased in the sense that some are under the original parameter, others over it, while ARS and all other methods denote exclusively positive shifts of the estimate versus the original parameter (see Figure 3).

The worse results of p-bootstrap with $\alpha = 4$ may be related to unfavorable features of the superset of 1,000 samples we used. Despite this large number, indeed, in Table 3 we realize the anomaly of this set by considering the mean of sample means and the value $\xi$ of the quantile level to which it corresponds. Namely, from the table we read this value for each $\alpha$; for instance, with $\alpha = 2$ and $\beta = 1$ the probability that the mean of 1,000 Gamma sample means is less than or equal to the actual value observed on the samples used in our experiment is 0.416. Through these data, we realize that the supersets used with $\alpha = 0.5$ and $\alpha = 4$ are somehow anomalous in that they correspond to very low quantile levels with respect to the others. The higher dispersion in the second case enhances this drawback. The analytic approximation strictly follows the pure numerical results, possibly with small unfavorable drifts as a toll for the computational benefits we will discuss later on. We do not report performances in the first two cases since we found the analytical approximation working well only for $\alpha \geq 2$, as highlighted in Figure 4. It seems to be due to the unsuitability

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi$</td>
<td>0.064</td>
<td>0.371</td>
<td>0.416</td>
<td>0.829</td>
<td>0.125</td>
<td>0.599</td>
</tr>
</tbody>
</table>

Table 3: Quantile level $\xi$ of the mean of the sample means observed on the 1,000 samples used to compute statistics in Table 2 for the different values of $\alpha$. 

URL: http://mc.manuscriptcentral.com/lssp E-mail: comstat@univmail.cis.mcmaster.ca
of the Gamma shape *per se*, since, for instance, in the picture with \( \alpha = 0.05 \) we left both \( \gamma_1 \) and \( \gamma_2 \) free in search of the best interpolation, which still proves unsuccessful.

As for estimates’ dispersion, we essentially lose w.r.t. ARS and some of the other methods as well. Indeed, in line with the statistical meaning of median \( \nu_X \) – namely it represents the parameter minimizing the sum of the absolute shifts, i.e.

\[
\nu_X = \min \{ \mathbb{E} [ |X - c|] \}
\]

where \( \mathbb{E} \) denotes the *expected value* operator – we use the median absolute deviation (MAD) as dispersion index which, with the above notation, reads:

\[
\text{MAD} = \nu_{|x - \nu_X|} = \text{med} [ |x - \nu_X|]
\]

where med denotes the *median value* operator. Now MAD is generally a lower dispersion indicator than the standard deviation \( \sigma \), where for a normal distribution we have in particular \( \text{MAD} = \sqrt{2/\pi} \sigma \). Thus, in spite of the small MAD values we may read in the table, we conclude that the population of our \( \alpha \) estimates is more dispersed than the analogous population obtained through ARS – while generally comparable with those of the other methods. We interpret this spreading as the counterpart of the estimator unbiasedness. To obtain this property, indeed, we do not rely on prior distributions, with a consequent broadening of the compatible parameter set.
Figure 4: Templates of $A$ distributions computed within Algorithmic Inference approach. Black curves: empirical CDFs drawn from a sample of $p$-bootstrap replicas; gray curves: approximate analytical CDFs computed from a subsample of the $p$-bootstrap replicas. The curve index $\alpha$ refers to the Gamma parameter with which the samples have been drawn.

As for computational loads, we compare $p$-bootstrap with the Bayesian procedures, since both base their estimates on a population of replicas. If we assume equivalent the costs of inverting either $h'_\alpha$ in (5) or the conditional distributions of $A$ given $\beta$ and $\{x_1, \ldots, x_m\}$ in Son and Oh (2006), the computational complexities of the procedures as a function of the sample size and number of replicas are equal. On the contrary, the analytic solution (11) for $F_A$ through the Gamma approximation with parameters (12) and (13) is two order faster. Then, since the reduction ratio of the number of replicas is 25, we may say that the hybrid procedure is one order faster.

In order to understand the influence of the prior in the parameter estimate, we delved into one sample with size 200 and $\alpha = 4$, causing performance differences in line with the above overall statistics, as shown in Table 4. In this table we report both Bayesian methods to contrast our results, where ARS is the direct competitor and RS helps identify a scaling of
Table 4: First and second order statistics, and confidence intervals of the parameters $A$ and $B$ for a reconstruction experiment with $m = 200$, $\alpha = 4$, and $\beta = 1$. Central value and Dispersion columns as in Table 2. Confidence Interval format: confidence level [lower bound, upper bound].

<table>
<thead>
<tr>
<th>Method</th>
<th>Central value</th>
<th>Dispersion</th>
<th>Confidence Interval</th>
<th>Central value</th>
<th>Dispersion</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>4.0362</td>
<td>0.3917</td>
<td>[0.99 [3.1038, 5.1351]]</td>
<td>0.99</td>
<td>[0.8072, 1.3628]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARS</td>
<td>4.038</td>
<td>0.3756</td>
<td>[0.99 [3.0931, 5.032]]</td>
<td>0.99</td>
<td>[0.8202, 1.3703]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI$_{p-boot}$</td>
<td>3.929</td>
<td>0.2661</td>
<td>[0.99 [2.9899, 5.0393]]</td>
<td>0.99</td>
<td>[0.8253, 1.4354]</td>
<td></td>
</tr>
<tr>
<td>AI$_{hybrid}$</td>
<td>3.923</td>
<td>0.3032</td>
<td>[0.99 [3.0134, 4.8676]]</td>
<td>0.99</td>
<td>[0.8462, 1.4175]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[1.0694 0.0757]</td>
<td>0.95</td>
<td>[0.8766, 1.319]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.90 [3.3249, 4.5836]]</td>
<td>0.90</td>
<td>[0.9052, 1.2743]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[1.0674 0.0894]</td>
<td>0.95</td>
<td>[0.8891, 1.329]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.90 [3.2972, 4.5411]]</td>
<td>0.90</td>
<td>[0.9125, 1.2835]</td>
<td></td>
</tr>
</tbody>
</table>

the results, and both Algorithmic Inference methods, referring to p-bootstrap as for statistic templates and the hybrid version for a cheaper variant. With all methods the statistics are collected from $n = 5,000$ replicas. Confidence interval extremes are exactly the $\delta/2$ and $1 - \delta/2$ quantiles, respectively, of the variously computed $A$ and $B$ distribution functions, where $1 - \delta$ is the confidence level. The comparison of the table values are as expected. In particular, we remark that the $\gamma_2$ coefficient of the analytical approximation is still based on 40 bootstrap replicas, thus realizing a speed up factor around 250. Instead, the exam of Figure 5 highlights a notably correlation between the replicas produced by ARS. Indeed, although the overall standard deviation is greater with p-bootstrap, the graph of the replica average by blocks of 100 consecutive replicas is more concentrated around the central value with our method than with the Bayesian one. Focusing on replicas distributions, in Figures 6
Figure 5: Courses of α replicas. (a) replicas generated by ARS; (b) replicas generated by p-bootstrap; (c) their averages by blocks of 100 consecutive replicas. Gray lines: ARS method; black lines: p-bootstrap method.

(a) and (b) we plot the 5,000 parameter estimates obtained solely with the p-bootstrap and ARS methods. We clearly recognize from the plots that distribution (17) inhibits high values of β that, *vice versa*, are acknowledged to be compatible with our method. This results in privileging high values of the α estimate (for instance in force of (18)) with the mentioned bias on their mean as a consequence.

In these figures we also draw confidence regions for the pair (α, β) that follow the thread of the parameter population (ARS regions narrower and shifted re p-bootstrap’s) and in any case look more suitable than the cartesian product of the confidence intervals reported in Table 4. To get these regions we use a standard peeling procedure (Liu et al., 1999; Apolloni et al., 2007b) on the bundle of Gamma CDFs obtained by plugging the estimated parameters in their general expression. Starting from the whole population of curves in Figures 6 (c) and (d) we iteratively visited the upper and lower bounds of the region they span, circularly removing the bordering curves in number of $\delta n/2$ lying below the envelope of the remaining ones and $\delta n/2$ curves above it, having denoted with $\delta$ the preset confidence level.

Performing a similar analysis on the problem of fitting the Gross and Clark sample with a Gamma distribution lets to draw the following conclusions. Table 5 denotes a great difference between the estimates of the fitting parameters. The analysis of their scatterplots denotes a definite unbalancing of ARS parameters that completely miss the left-uppermost region.
Figure 6: Distributions of Gamma parameters and related CDFs obtained with two inference approaches. Left column: Algorithmic Inference approach with p-bootstrap method; right column: Bayesian approach with ARS method. First row: parameters’ scatterplot contoured by 0.90 confidence regions; plain line: contouring through p-bootstrap; dashed line: contouring through ARS. Second row: bundle of Gamma CDFs with the parameters plotted in the first row. Gray filled areas: 0.90 confidence regions, with the column method, corresponding to the companion regions in the first row.
<table>
<thead>
<tr>
<th>Method</th>
<th>Central value</th>
<th>Dispersion</th>
<th>Confidence Interval</th>
<th>Central value</th>
<th>Dispersion</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>9.2120</td>
<td>2.6430</td>
<td>0.99 [3.7912, 17.1581]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95 [4.7248, 14.9754]</td>
<td>0.90 [5.2521, 13.8601]</td>
<td>13.5246</td>
<td>4.4529</td>
<td>0.95 [7.4934, 24.8940]</td>
<td>0.90 [8.0750, 22.0373]</td>
</tr>
<tr>
<td></td>
<td>0.95 [4.6920, 15.3352]</td>
<td>0.90 [5.3633, 14.0487]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI&lt;sub&gt;p-boost&lt;/sub&gt;</td>
<td>13.4671</td>
<td>4.5923</td>
<td>0.99 [3.7973, 43.6594]</td>
<td>8.4424</td>
<td>2.8444</td>
<td>0.95 [3.2778, 23.7440]</td>
</tr>
<tr>
<td></td>
<td>0.95 [4.8630, 34.1774]</td>
<td>0.90 [5.6707, 29.4574]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95 [4.6995, 29.651]</td>
<td>0.90 [5.4986, 26.188]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95 [4.6995, 29.651]</td>
<td>0.90 [5.4986, 26.188]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Same statistics as in Table 4 collected on G&C dataset.

of the parabolic-like distribution. This imbalance, which is opposite to the one studied before, has another explanation. The histogram in Figure 7 clearly shows that the Gross and Clark sample has a very low likelihood when it is attributed to a Gamma distribution. Actually, we do compute the ML estimate, for instance with the approximation proposed in Thom (1958), of these parameters. However, if we generate new samples exactly with these parameters and compute their likelihood, we realize that the likelihood of Gross & Clark sample represents around the fifth quantile of the likelihood distribution, thus getting a far location from the core in the histogram shown in the figure. In other words, as a sample of a Gamma distribution, it is a rather anomalous sample. Thus the p-bootstrap method gives rise to a broad parameter distribution that maintains, however, a due symmetry, as shown in Figure 8 (a) and (b). The plots of the distribution bundle denote a wider sheaf and confidence region with our method (see Figure 8 (c) and (d)). But, once again, we expect this to be rewarded by the unbiasedness of estimates, which in this case are definitely far from those obtained with conventional methods. As a final remark, the numerical results are taken on the basis of 10,000 parameter replicas, while hybrid method used only 200 replicas,
Figure 7: Histogram of the negative log-likelihood of samples drawn from a Gamma distribution with parameters given by the ML\textsuperscript{Thom} estimates from G&C dataset. Thick bar: likelihood of G&C dataset with this parameterization.

entailing a great saving of running time. Moreover, the $\gamma_1$ expression (16) derives from data interpolation when $m$ uniformly ranges in [20, 300], $\alpha$ in [2, 6], and $\beta = 1$. Thus this case highlights a great extendability of the interpolated function to a range not covered by the training sample, even though closer results to full bootstrap method’s may be expected with a more focused interpolation.

4. CONCLUSIONS

Facing the non trivial inference problem of estimating the two parameters of a Gamma distribution, we propose an estimator exhibiting the distinguishing feature of being unbiased, while a large set of reputed estimators does not enjoy the same property, with a further drawback of proving less accurate in a set of reconstruction problems. To obtain this result we adopted a different statistical framework showing the comparative benefit w.r.t. the Bayesian approach of not needing a prior parameter distribution. *Per se* contrasting the two approaches may appear no different from the Lilliputian contest described by Jonathan Swift in Gulliver Travels *re the way of breaking eggs: upon the larger or smaller end*. In place of the *six rebellions raised on that account* by Lilliputians, we just claim that missing prior distributions may improve the statistical properties of the estimators. This reopens a major dispute fought in the mid 20th century about the interpretation of parameter variability.
(Zabell, 1992) in terms of fiducial distribution (Fisher, 1956), structural probabilities (Fraser, 1966), priors/posteriors (Ramsey, 1925), and so on. From an epistemologic viewpoint, this entailed a companion dispute as to the way of modeling probability: as a physical feature of the phenomena to be described through random variables or a tool for synthesizing data about a phenomenon? For a long time, the former model won out for all practical purposes. However the results we obtain in this way may prove narrow, in that almost harmless when the candidate models to explain data cannot involve Gaussian probability distributions or asymptotic limit theorems (Meeker and Escobar, 1995). Bayes approach offers a way to escape this drawback, by considering a priori parameter distribution laws – sometimes linked to the features of the subtending phenomena (Florens et al., 1990; Berger, 1999), other times just as an analytical artifact (the uninformative priors (Jaynes, 1968)) – which are modified

![Figure 8: Same pictures and notation as in Figure 6 when the sample is G&C dataset.](image-url)
by the information content of the observed data. Very often this proves to be only the shift of the hard part of the inference problem one step back, i.e. to the answer to the question: Who tells us the priors?

We propose here and in some companion papers in the frame of computational learning (Apolloni et al., 2007c, 2008b,a, 2007b; Apolloni and Bassis, 2008; Apolloni et al., 2006, 2008c) to return to the original thought of Fisher of considering the model parameters as random variables in light of the observed data (Fisher, 1956). With this perspective, he defines a fiducial distribution law of parameters of a given random variable that he deduces from a sample of its realizations. Through this law he computes, for instance “the probability that $\mu$ (mean of a Gaussian variable – our note) is less than any assigned value, or the probability that it lies between any assigned values, or, in short, its probability distribution, in the light of the sample observed” (Fisher, 1935). The Achilles heel of Fisher’s approach lies in the joint distribution of more than one parameter (Zabell, 1992). In this paper we show that this is not longer true with our revisitation, thanks to the wide availability of computational power. Through (possibly heavy) numerical procedures indeed, we may invert master equations like (5) to reverse the probability masses we jointly distribute on the variable realizations for set parameters – a well assessed attitude – into the joint probability distribution of the parameters compatible with the observed data. The compatibility notion is very close to the fiducial probability or the structural probability proposed respectively by Fisher (1956) and Fraser (1961). In this paper we both supply algorithmic procedures for its computation and stress its properties on a severe reconstruction problem.

References


URL: http://mc.manuscriptcentral.com/lssp E-mail: comstat@univmail.cis.mcmaster.ca


Response to referee comments in the paper:

ID LSSP-2009-0057, Algorithmic Inference of Two-Parameter Gamma Distribution

*General

We have revised the paper according to the suggestions of the referee, whom we thank sincerely for his consideration.

Specifically, the changes we have made on the basis of the referee’s comments are as follows:

C(omment). Different seeds than uniform variable; the rationale of “seed” notion.

A(nswer). On page 4, line 10 and ff. we mention that the standard Gaussian variable is very often used as a seed in the practice of variable simulation. We also explain the benefit of introducing the notion of seed as a link between parameter and variable randomness.

C. s_3 monotony.

A. On page 7, line -11 we clarify that we numerically realized this property. Actually, we already remark on page 5, line 11, that (s_1,s_3) constitute a different representation of the jointly sufficient statistics (s_1,s_2), but we have no direct analytical proof of the s_3 monotony with \alpha.