IntechOpen

Statistical Methodologies

Edited by Jan Peter Hessling

Statistical Methodologies

Edited by Jan Peter Hessling

IntechOpen

Supporting open minds since 2005

Notice
Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

We are IntechOpen,
the world's leading publisher of Open Access books
Built by scientists, for scientists

4,600+
Open access books available

120,000+
International authors and editors

135M+
Downloads

151
Countries delivered to

Our authors are among the

Top 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS

BOOK CITATION INDEX

INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Meet the editor

Jan Peter Hessling earned his PhD in Theoretical Physics from Chalmers University of Technology, Gothenburg, Sweden, in 1996, and his MSc in Physics from the University of Massachusetts, USA, in 1991. Since then he has been devoted to the development of novel mathematical concepts, recently focusing on modeling uncertainty and dynamic analysis. Peter is the original proposer of *Deterministic Sampling* for uncertainty quantification and model calibration, and *Dynamic Metrology* for analysis of dynamic measurements utilizing custom digital filtering. Peter has authored three book chapters for InTech and about 20 journal articles as leading author; his current Google h-index = 12. Since 2016 his concepts have been further developed in the privately held company Kapernicus AB, dedicated to applied mathematical R&D, with services offered worldwide.

Contents

Preface

"Why Most Published Research Findings Are False"

- John P.A. Ioannidis

The recalcitrant and alarming title of Ioannidis' research article from 2005 (quoted above) addressed a growing concern that the current practice of statistics to a significant extent could be flawed. In his book *Willful Ignorance*: *The Mismeasure of Uncertainty* (2009), Herbert Weisberg pointed out the nearly abyssal difference in perspectives of quantitative and qualitative studies and reflected on the historical evolution of statistics, displeased by the fact that practical scientists nowadays systematically avoid discussing the meaning of probability, despite its being intensively debated for over two centuries. These claims of doubt were accompanied by Alex Reinhard, who in his book *Statistics Done Wrong: The Woefully Complete Guide* (2015) discussed a number of questionable ways statistical analyses of today are performed.

The famous statistician R.A. Fisher stated in the early 20th century that: "*The tendency of modern scientific teaching is to neglect the great books, to lay far too much stress upon relatively unimportant modern work, and to present masses of detail of doubtful truth and questionable weight in such a way as to obscure principles.*" In the modern society flooded not only by massive unleashed computational power but also by data-based artificial intelligence and unprecedented means to craft stunning graphics, it is not at all surprising that people at large are struggling to distinguish facts from fake. Fisher's concern from the 1930s suddenly appears more alive than ever before. The need to re-evaluate statistical practices thus steadily seem to grow the more they are utilized.

These observations suggest current statistical methodologies may have gone too far into an engineering practice with minimal questioning. As an example, multiple tests are often launched in scientific studies to enhance the likelihood of finding a significant result, although this should be impossible …

In the modern age pervaded by data in various forms, statistical processing affects almost everyone, whether we like it or not. Internet-based analyses operate on sample sets almost as big as the populations they are drawn from. Behavioral, political, and customer surveys are frequently made. Despite this, the basic principles and practices have remained remarkably intact over the years. In stark contrast stands the rapid evolution of computer programming methodologies and development of convenient, flexible, high-level, and broadly accessible software tools for statistical analysis and presentation, like Python, R, Matlab/Octave, etc. To fully utilize their potential rather than be distracted and paralyzed by their breathtaking performance, good statistical methodologies and practices are of greater value than ever before. This context provided the basic motivation for this book.

A selection of various aspects of statistical methodologies are here presented by independent authors. The chapters are not meant to be exhaustive or representative, even though each contribution is self-contained and complete within the task

addressed. The texts are primarily meant to motivate further reading. Being overviews, many details are left out. The reader is therefore advised to make extended use of the list of references presented at the end of each chapter.

The introductory chapter discusses possible consequences of our willful ignorance, or what is often assumed but seldom known. The main part of the book starts with methods to analyze sensitive questions and how reliable unbiased claims can be extracted. Two-phased stratified sampling is then studied for various estimators. *To facilitate search of prior art when authoring and examining patents, a methodology based on machine learning for analyzing and structuring patents is discussed thereafter. Density estimation in inventory control systems*, a comparison of maximum likelihood and Bayesian estimation for the Erlang distribution follows, before the asymptotic normality of Hill's estimator is addressed. The book concludes with a study of the Jackknife resampling method.

Current practice has indeed been questioned numerous times in recent years, even though mathematical statistics for a very long time has been widely applied in medicine, social studies, science, and technology. Hopefully, this book will provide guidance into some aspects of statistical data collection and processing for the future.

Jan Peter Hessling
Kapernicus AB,
Gothenburg, Sweden

Introductory Chapter: Ramifications of Incomplete Knowledge

Jan Peter Hessling

"*Facts do not cease to exist because they are ignored.*"

1. Background

Mathematical statistics has long been widely practiced in many fields of science [1]. Nevertheless, statistical methods have remained remarkably intact ever since the pioneering work [2] of R.A. Fisher and his contemporary scientists early in the twentieth century. Recently however, it has been claimed that most scientific results are wrong [3], due to malpractice of statistical methods. Errors of that kind are not caused by imperfect methodology but rather, reflect lack of understanding and proper interpretation.

In this introductory chapter, a different cause of errors is addressed—the ubiquitous practice of willful ignorance (WI) [4]. Usually it is applied with intent to remedy lack of knowledge and simplify or merely enable application of established statistical methods. Virtually all statistical approaches require complete statistical knowledge at some stage. In practice though, that can hardly ever be established. For instance, Bayes estimation relies upon prior knowledge. Any equal a priori probability assumption ("uninformed prior") does hardly disguise some facts are not known, which may be grossly deceiving. Uniform distribution is a specific assumption like any other. Willful ignorance of that kind must not be confused with knowledge to which we associate some degree of confidence. It may be better to explore rather than ignore consequences of what is not known at all. That will require novel perspectives on how mathematical statistics is practiced, which is the scope of this book.

2. Ambiguity

Incomplete knowledge implies that obtained results may not be unique. That is, results may be ambiguous. Ambiguity de facto means the uncertainty associated with any estimated quantity itself is uncertain. We may adopt a probabilistic view and classify ambiguity as epistemic uncertainty. Ambiguity will here refer to lack of knowledge typically substituted with willful ignorance. Alternatives propelled by different types of willful ignorance can thus be explored to assess ambiguity.

A most powerful source of ambiguity is dependencies. Independence is perhaps the most claimed but often the least discussed presumption. Throwing dices or growing crops, as typically studied by the founders of statistics, independence indeed seems plausible. In all the complexity of modern technology of today however, it is anything but evident observations are independent. For instance,

meteorological radar observations may share sources of errors, meaning recorded data will be statistically dependent. A problem may then arise if our analysis makes use of, e.g., the maximum likelihood method which utilizes the entire covariance matrix. Most of its entries, all covariances between pairs of observations, are usually not known but bluntly set to zero to enable evaluation. This willful ignorance has the drastic consequence of extinguishing ambiguity and, as will be shown, minimizing the resulting uncertainty. Elementary considerations should provide the valuable insight that even exceedingly small covariances may substantially influence the result: the number of covariance elements is $n(n-1)/2 \approx n^2/2$, while there are only n variances, for n observations. The number of covariance elements is hence $n/2$ times larger than that of variance. Each element being small is thus not a good enough argument to ignore the collection of all covariance elements.

Various attempts have been made to avoid willful ignorance. The method of maximum entropy [5] focuses on the consequences of improper assignments of unknown statistical information. Covariance intersection [6] fuses observations conservatively to a pair of uncorrelated observations with variance $\max[\mathrm{var}(\theta)]$. This approach explores ambiguity along the general principles suggested here, considering all possible values of covariance. Complementing the obtained maximum variance with the least possible variance $\min[\mathrm{var}(\theta)]$ would render an ambiguity interval, $A = \max[\mathrm{var}(\theta)] - \min[\mathrm{var}(\theta)]$, different but similar to confidence intervals.

Repeating any statistical analysis with various kinds of willful ignorance [on its input], the ambiguity (A) [of its output] can be assessed. Some WI will give large, while others will yield small resulting uncertainty, not necessarily the maximum and minimum, as it is difficult to imagine all possible kinds of WI. Any specific WI will more or less reduce or quench the uncertainty from its maximum. Identifying a model from calibration data H_{CAL} and then letting the so-obtained model predict the same data H_{PRD}, any chosen willful ignorance of $\mathrm{cov}(H_{\mathrm{CAL}})$ will quench the calibration uncertainty from the maximum over all choices, $\mathrm{var}(H_{\mathrm{PRD}}) \leq \max[\mathrm{var}(H_{\mathrm{PRD}})] \leq \mathrm{var}(H_{\mathrm{CAL}})$. Studying uncertainty quenching through $\mathrm{var}(H_{\mathrm{CAL}}) - \mathrm{var}(H_{\mathrm{PRD}})$ will indicate possible ramifications of our lack of knowledge $\mathrm{var}(H_{\mathrm{PRD}}) \leq \max[\mathrm{var}(H_{\mathrm{PRD}})]$ but also the implicit knowledge $\max[\mathrm{var}(H_{\mathrm{PRD}})] \leq \mathrm{var}(H_{\mathrm{CAL}})$ contained in the structure of the model. Most importantly, such studies will guide us to the least harmful choice of willful ignorance. The analysis is similar in style but different to the method of maximum entropy and covariance intersection. An example is given below.

3. Illustration of uncertainty quenching

Assume we would like to study the evolution of a field over two spatial coordinates, using a model composed of a set of differential equations. The field could refer to meteorology and describe current observations of air pressure or humidity. The initial state may be expanded in the set of basis functions of the appropriate operator, similar to forecasting in numerical weather prediction (NWP) [7]. The basis functions could be thought of as the eigensolutions of a linear operator, which propagates one meteorological state, from one day to another. Neither the interpretation of the field nor the field itself matters for the discussion here. Rather, it is how the uncertainty of the initial state is represented as uncertainty of the distributed eigensolutions of the NWP propagator. This representation will determine the uncertainty of any subsequent forecast, reflecting the past experience in future confidence of predicting the weather. If the forecast uncertainty is lower than our current knowledge reflects, we may falsely reject, e.g., the possibility of experiencing major thunderstorms. In the eye of sailors planning their journey, the forecast

uncertainty is the indisputable decision-maker. Studying the uncertainty quenching $\text{var}(H_{\text{CAL}}) - \text{var}(H_{\text{NWP}})$, the ambiguity regarding the usually unknown but nevertheless required covariances $\text{cov}\left(H_{\text{CAL}}(x_i, y_i), H_{\text{CAL}}\left(x_j, y_j\right)\right)$ can be assessed. Then by expanding the uncertainty conservatively, serious events like major thunderstorms may be properly recovered.

To enable illustrations, let the eigenstates of the NWP operator of order n [for not known reasons] be multiplicative separable in time t as well as in spatial coordinates x, y, with eigenstates described by orthogonal Legendre polynomials up to order n:

$$H_{\text{NWP}}^{(n)}(x, y, t) \equiv \sum_{j,\,k=0}^{n} \theta_{\underbrace{j + k(n+1) + 1}_{\equiv r}}(t) \cdot P_k(y) P_j(x), \qquad (1)$$

where the NWP operator propagates the coefficients $\theta_{j+k(n+1)+1}(t)$ in time. Only the representation of $\text{cov}(H_{\text{CAL}}(x, y, t = 0))$, the covariance of the measured initial state at $t = 0$, is of interest here. Discretizing over m domains in both directions x, y followed by sequential scanning over x_p for each y_q, $p, q = 1, 2, \ldots m + 1$, the model is written in standard affine vector $(\overline{\theta}, \overline{K})$ form:

$$H_{\text{NWP}}^{(n)}\left(x_p, y_q, 0\right) \equiv \sum_{r=1}^{(n+1)^2} \theta_r(t = 0) K_r\left(x_p + (m+1)y_q\right) = \overline{\theta}^T \overline{K}. \qquad (2)$$

Without any supplementary information, the variance of the initial measurement should be completely represented by the variance of the initial model state, i.e., $\text{var}(H_{\text{NWP}}(x, y, 0)) = \text{var}(H_{\text{CAL}}(x, y, 0))$. The question is whether this holds, and if it does not, to which extent can we minimize the discrepancy with WI?

Assuming normal distributed measurement noise, the maximum likelihood method [8] yields the parameter covariance given by Eq. (3), which is propagated to uncertainty of the best predictions according to Eq. (4):

$$\text{cov}(\theta^*) = \left[K \text{cov}(H_{\text{CAL}}) K^T\right]^{-1}, \qquad (3)$$

$$\text{cov}(H_{\text{NWP}}) = K^T \text{cov}(\theta^*) K \qquad (4)$$

Combining these relations, the degree of completeness of the representation of uncertainty by the model can be studied:

$$\text{var}(H_{\text{NWP}}) = \text{diag}\left\{K^T \left[K \text{cov}(H_{\text{CAL}}) K^T\right]^{-1} K\right\} \cong \text{var}(H_{\text{CAL}}), \qquad (5)$$

where $\cong$ indicates the addressed equality in the absence of uncertainty quenching or maximal propagation of uncertainty from observations to model. Equality can never be achieved though, since the number of degrees of freedom (NDOF) of prediction is drastically lower than that of calibration data. For typical models and data, the two NDOFs usually differ by an order of 10 or more. A large ratio is actually required to provide sufficient redundancy. As seen in **Figure 1**, the uncertainty is normally reduced to a small fraction, with substantial uncertainty quenching.

It should be emphasized that stating independence is fundamentally different than stating the degree of dependence which is unknown. These statements in fact oppose each other, since independence maximizes the available amount of information. Indeed, the Fisher information matrix [9]

$$\frac{1}{\text{CRLB}} = F\left(H_{\text{CAL}, i}, H_{\text{CAL}, j}\right) = -\left\langle \frac{\partial^2 \ln(p)}{\partial H_{\text{CAL}, i} \partial H_{\text{CAL}, j}} \right\rangle = \begin{cases} F_1, & m \text{ dep.} \\ m F_1, & m \text{ indep.} \end{cases} \qquad (6)$$
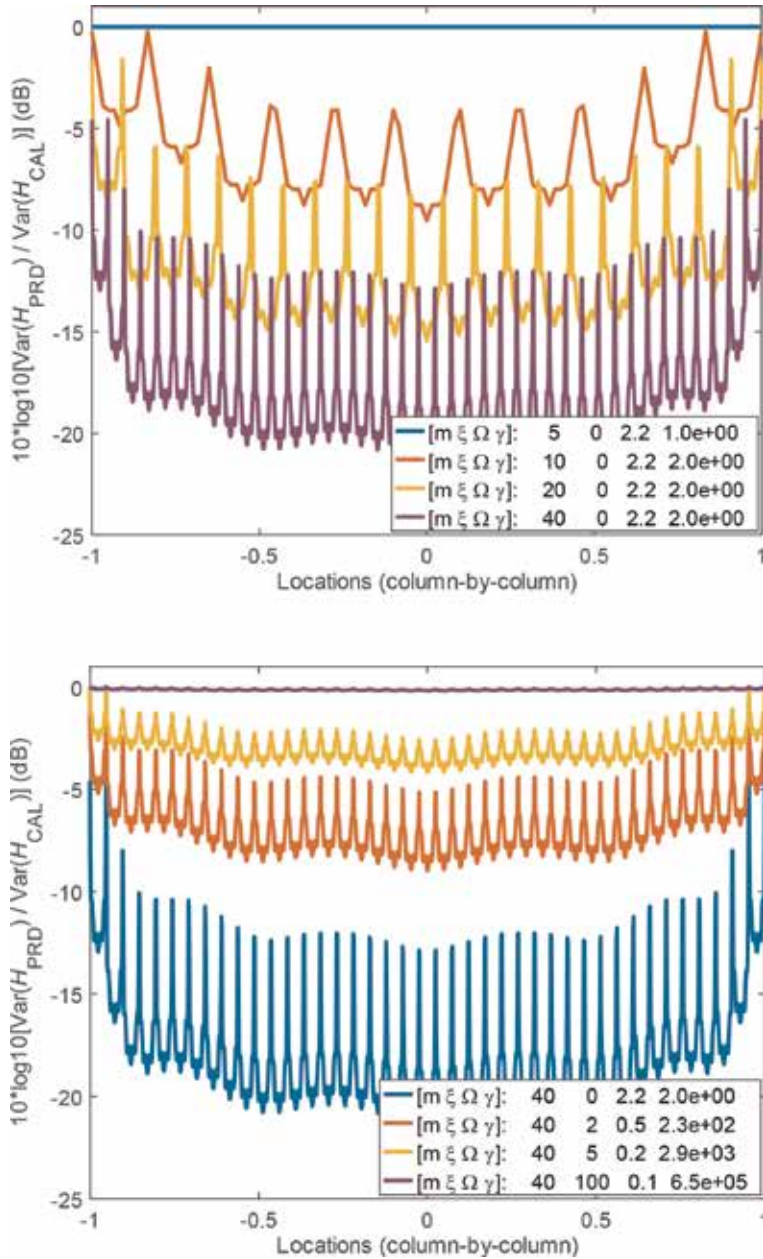
Figure 1.
Uncertainty quenching or excessive reduction of uncertainty due to willful ignorance. Dependence on resolution m (top) and correlation length ξ (Eq. (7)) (bottom) of the calibration data. The legend includes the Mahalanobis canonical distance Ω (Eq. (11)) and the ratio γ between the largest and the smallest eigenvalues of residual variance (Eq. (9)).

is additive as m independent observations are collected, since the joint probability distribution p in that case is multiplicative separable. For dependent observations though, no information is added at all, since $F\left(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j}\right)$ then remains the same as for only one observation (F_1). Aggregating m observations, the minimum variance of any estimator set by the Cramer-Rao lower bound (CRLB) [9] thus decreases with a factor $1/m$ in the case where errors are independent (Eq. (6), first row), while it remains the same if they are completely dependent (Eq. (6), second row). Independence is thus the worst possible choice of WI, as it builds

confidence without knowledge. WI should minimize rather than maximize the information. That is indeed the principle utilized in the method of maximum entropy [5].

Uncertainty is lost for obvious reasons. The question is how much and for what reason. Since the model cannot represent an arbitrary response, it can neither represent an arbitrary variability. This restriction constitutes the very meaning of a "model." This makes it important to describe the covariance of observations accurately—inappropriate WI may quench uncertainty dramatically.

The additional information represented by the structure of the model could be denoted by the model innovation. It is strongly affected by WI attributed to observations. With increasing resolution m, the model innovation grows as the information contained in observations is maximized with an assumption of independence. Indeed, the prediction variance is quenched in agreement with the CRLB, as seen in **Figure 1** (top).

If WI of observation covariance instead resembles what the model is able to represent, the model innovation will be the least. Instead of assuming independent observations, introduce a finite long correlation length\ksi:

$$\text{corr}\big(H_{\text{CAL},i}, H_{\text{CAL},j}\big) = \frac{\text{cov}\big(H_{\text{CAL},i}, H_{\text{CAL},j}\big)}{\sqrt{\text{var}(H_{\text{CAL},i})\text{var}\big(H_{\text{CAL},j}\big)}} = \exp\big[-\big|r_i - r_j\big|/\xi\big] \qquad (7)$$

Increasing the correlation length\ksi from zero as in **Figure 1** (bottom), the model innovation decreases, and the variance of the prediction $\text{var}(H_{\text{NWP}})$ is almost fully restored to the original variance of the observation $\text{var}(H_{\text{CAL}})$. The model will then not improve our knowledge of the current weather situation but enable prediction to a later time with comparable trust. Summarizing **Figure 1**, our WI of observation covariance and resolution m strongly influence our claimed precision $\text{var}(H_{\text{NWP}})$ of predictions.

It is a different matter if the model is consistent with the observations it was identified from. Model consistency is usually assessed with a statistical residual analysis. In conventional system identification (CSI) [10], the hypothesis is that the [deterministic] model fully explains the observations. Due to sampling variance of the finite uncertain calibration data though, the best estimate of its parameters will be uncertain. The residual analysis explores if the residual is consistent with the sampling uncertainty of the calibration data but without uncertainty associated with the model.

This conjecture of a model without error whatsoever in CSI is questionable. In practice, no model is completely without error. Rather, a finite uncertainty of the model could be regarded as inherited from mismatch to calibration data. If so, the model merely provides a convenient but to a quantifiable degree imperfect basis for expressing uncertain calibration data. The model is utilized to "passively transform" rather than "actively explain" observations to another unknown situation of interest. That intent is typical in, e.g., weather forecasting and product development. Furthermore, the uncertainty of calibration data can often be assessed from the setup of the calibration experiment. In CSI correlation functions are evaluated from a single residual vector, enforcing homoscedasticity and independence of observations. WI of this kind enables the statistical analysis of the residual but often find little support.

The alternative view on model calibration proposed here is that the identified model, composed of its form or structure, parameters, and uncertainty, represents the uncertain calibration data. Model results can thus substitute our observations, to the degree various aspects of the model and observations are consistent. Any given

residual is one realization and should relate to its expected variability, with respect to the uncertainties of both the model and the observations it was identified from.

The Mahalanobis distance [6] can be utilized to measure the relative distance between observations and model output, which constitutes the residual ρ:

$$M = \rho^T \mathrm{cov}^{-1}(\rho)\rho, \quad \rho \equiv H_{\mathrm{CAL}} - H_{\mathrm{PRD}}. \tag{8}$$

The residual covariance matrix defines its principal variations with typical magnitudes λ_j:

$$\mathrm{cov}(\rho) = \mathrm{cov}(H_{\mathrm{CAL}}) - 2\mathrm{cov}(H_{\mathrm{CAL}}, H_{\mathrm{PRD}}) + \mathrm{cov}(H_{\mathrm{PRD}}) = U\Lambda^2 U^T,$$
$$\Lambda_{ij} \equiv \delta_{ij}\lambda_j, \quad UU^T = I, \quad \delta_{jj} = 1, \quad \delta_{i \neq j} = 0. \tag{9}$$

The evaluation of $\mathrm{cov}(H_{\mathrm{CAL}}, H_{\mathrm{PRD}})$ in Eq. (9) is challenging, since H_{PRD} has a complicated relation to its "role model" H_{CAL} set by the identification. To simplify, it is set to zero below.

Extracting matrices U from Eq. (9) in Eq. (8), squared deviations are compared to variances. The Mahalanobis distance then transforms into a relative Euclidean norm of the residual in its own space of uncorrelated variations:

$$M \equiv \tilde{\rho} T \Lambda^{-2} \tilde{\rho} = \sum_j \frac{\left|\tilde{\rho}_j\right|^2}{\mathrm{var}\left(\tilde{\rho}_j\right)}, \quad \tilde{\rho} \equiv U\rho, \tag{10}$$

where $\tilde{\rho}_j$ is the projection of the residual on its principal vector $U_{\cdot,j}$ of variation, while the eigenvalue $\lambda_j \equiv \{\Lambda\}_{jj}$ expresses its typical magnitude of variation. For a small eigenvalue λ_j, observing even a moderate projection $U_{j,\cdot}\rho$ is statistically unlikely and thus strongly violates any model.

To maximize the consistency, in the sense of minimizing the Mahalanobis distance, the variance $\mathrm{var}\left(\tilde{\rho}_j\right)$ of principal residual variations should be maximized. Without addressing any specific residual, maximize what could be defined the Mahalanobis canonical distance:

$$\Omega \equiv \sqrt{\sum_j \mathrm{var}\left(\tilde{\rho}_j\right)} = \sqrt{\sum_j \lambda_j^2}. \tag{11}$$

Minimizing the Fisher information matrix under assumption of normality addresses the covariance $\mathrm{cov}\left(H_{\mathrm{CAL},i}, H_{\mathrm{CAL},j}\right)$ of observations. Minimizing the Mahalanobis canonical distance Ω considers also the covariance $\mathrm{cov}\left(H_{\mathrm{PRD},i}, H_{\mathrm{PRD},j}\right)$ of the model residual as well as the cross covariance $\mathrm{cov}\left(H_{\mathrm{CAL},i}, H_{\mathrm{PRD},j}\right)$, which reflects the model innovation. Hence, willful ignorance for model identification should minimize the Mahalanobis canonical distance rather than the Fisher information matrix, as the former but not the latter also accounts for the innovation of the model structure. The intent is to educate the model to produce the most conservative results.

In practice, no residual projection $\tilde{\rho} = U\rho$ is usually negligible. Thus, the likelihood of rejecting any model, considering correlations, increases dramatically if exceedingly small eigenvalues are obtained. For that reason it is wise to check the ratio $\gamma \equiv \max_j (\lambda_j) / \min_j (\lambda_j)$ between the largest and smallest eigenvalues of the residual covariance (Eq. (9)). If γ is large, the model is expected to fail with respect

to correlations of the residual. The model may very well be consistent with respect to the variance but rarely with respect to covariance of its output. However, ignoring correlations and only focusing on the magnitude of variations of calibration data, i.e., $\text{var}(H_{\text{CAL}})$, which is the standard practice [10], is completely different. Then, the belief in the model is perfect and the only limitation of also making perfect forecasts is the finiteness of a random sample of observations. In case of homoscedasticity, $\gamma = 1$.

A potential conflict is inevitable for exceedingly high ratios γ. Indeed, as seen in **Figure 1** (bottom) for increasing correlation lengths ξ, the Mahalanobis distance Ω decreases, while the ratio γ rapidly increases. Thus as observation variance is recovered, the requirement to ignore prediction covariance rapidly grows.

4. A quest for better practice of willful ignorance

> *"The first principle is that you must not fool yourself and you are the easiest person to fool"* [11].

Current practice of willful ignorance sometimes makes statistics an art of self-delusion [3]. Consequences of applied WI are rarely explored, as only one proposition normally is made without further ado.

Distinguishing what is not known from what is assumed is of paramount importance. Not known to any degree should mean that all possibilities that can be imagined also ought to be considered. Otherwise obtained results only exemplify what the most appropriate answer may be, without any indication of the largest possible deviation.

Our knowledge is almost never complete. Virtually all existing statistical methods nevertheless require precisely that. Until alternative methodologies exist, WI must fill the gap between what is actually known and what must be known. As illustrated, the consequences of different WI may vary dramatically. Therefore we should select and tweak WI carefully. WI should not relate to our unconfirmed belief, but rather address its consequences.

The proposal of a quantifiable ambiguity proposed here suggests how ramifications of incomplete knowledge might be mitigated with carefully chosen WI: explore all kinds of ignorance that can be imagined. Analyze and collect obtained results in ambiguity intervals, similar to confidence intervals. Another option is to focus on the worst case in a conservative manner. The method of covariance intersection is one example of how that can be exercised. The principle of maximum entropy provides means to maximize the residual uncertainty, to add the least possible amount of information. Minimizing the Fisher information for observations and the Mahalanobis distance for model identification as proposed here is still another kind of conservatism. These methods tackle unknown information with WI and explore its consequences. Finding the most proper WI is indeed nontrivial and calls for genuinely novel approaches.

Current practice of statistics utilizes WI in many ways, but the specific choice is rarely discussed in depth. One reason could be that statistics was developed in an entirely different context than practiced today, which is rarely acknowledged and probably not fully comprehended. To exemplify, recall that Fisher's [2] original interpretation of "never" as a finite probability of 5% was just a humble proposal. He urged his readers to adjust "never" to the current context, a piece of advice almost never followed today.

Perhaps the reported breakdown of statistics methodologies [3, 4] is due to neglect of ambiguity, driven by a strong tradition of uncritical application of WI.

Could this be caused by lack of awareness of its potentially dramatic consequences? Ignorance of limitations of contemporary state-of-the-art methods is hardly new [12]. Ambiguity indeed sets a meta-perspective on statistical analysis that cannot be avoided and thus needs further exploration.

Author details

Jan Peter Hessling
Kapernicus AB, Hallingsjo, Sweden

*Address all correspondence to: peter@kapernicus.com

IntechOpen

References

[1] Huxley A. Proper Studies. London: Chatto and Windus; 1927

[2] Bennett JH, Fisher RA. Statistical methods, experimental design, and scientific inference. Fisher. New York: Oxford University Press; 1995

[3] Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;**2**(8):e124. DOI: 10.1371/journal.pmed.0020124

[4] Weisberg HI. Willful Ignorance: The Mismeasure of Uncertainty. Hoboken, New Jersey: John Wiley & Sons; 2014. ISBN: 978-0470890448. ISBN: 0470890444

[5] Jaynes ET. Information theory and statistical mechanics. Physical Review. 1957;**106**(4):620

[6] Uhlmann JK. Covariance consistency methods for fault-tolerant distributed data fusion. Information Fusion. 2003;**4**(3):201-215

[7] Kalnay E. Atmospheric Modeling, Data Assimilation and Predictability. New York: Cambridge University Press; 2003. ISBN: 978-0-521-79179-3. ISBN: 978-0-521-79629-3

[8] Hessling JP. Identification of complex models. SIAM/ASA Journal on Uncertainty Quantification. 2014;**2**(1): 717-744

[9] Kay SM. Fundamentals of Statistical Signal Processing, Estimation Theory. Vol. 1. Upper Saddle River, New Jersey: Prentice Hall PTR; 1993

[10] Ljung L. System Identification—Theory for the User. Englewood Cliffs, NJ: Prentice-Hall; 1987. 519 p. ISBN 0-13-881640

[11] Feynman RP. Cargo Cult Science: Some remarks on science pseudoscience, and learning how to not fool yourself. Caltech's 1974 commencement address. Engineering and Science. 1974;10-13

[12] Fisher RA. On the mathematical foundations of theoretical studies. Philosophical Transactions of the Royal Society A. 1922;**222**:309-368. Reproduction with Author's note: https://web.archive.org/web/20051213222222; http:/www.library.adelaide.edu.au/digitised/fisher/18pt1.pdf

Surveying Sensitive Topics with Indirect Questioning

Evrim Oral

Abstract

Data reliability is a common concern especially when asking about sensitive topics such as sexual misconduct, domestic violence, or drug and alcohol abuse. Sensitive topics might cause refusals in surveys due to privacy concerns of the subjects. Unit nonresponse occurs when sampled subjects fail to participate in a study; item nonresponse occurs when sampled subjects do not respond to certain survey questions. Unit nonresponse reduces sample size and study power; it might also increase bias. Respondents, on the other hand, might answer the sensitive questions in a manner that will be viewed favorably by others instead of answering truthfully. Social desirability bias (SDB) has long been recognized as a serious problem in surveying sensitive topics. Various indirect questioning methods have been developed to reduce SDB and increase data reliability, one of them being the randomized response technique (RRT). In this chapter, we will review some of the important indirect questioning techniques proposed for binary responses, with a special focus on RRTs. We will discuss the advantages and disadvantages of some of the indirect questioning techniques and describe some of the recent novel methods.

Keywords: social desirability bias, unmatched count technique, network scale-up technique, nonrandomized response technique, randomized response technique

1. Introduction: surveying sensitive topics

Data reliability is a common concern across all studies that use surveys, but more so while asking sensitive questions. Sensitive questions include sacred, private, or potentially exposing information that could be incriminating or discriminating for a respondent, or for the social group that is represented by the respondent [1]. For example, in studies which evaluate exposure to HIV infection, respondents are often asked sensitive questions regarding their opposite- or same-sex sexual practices. As another example, in studies which aim to assess substance use and abuse, respondents might suppress disclosure of their drug and alcohol misuse to avoid embarrassment or potentially harmful/unwanted consequences. Estimating the prevalence of such sensitive attributes is particularly important for health care researchers to build scientific knowledge, create necessary public health interventions, and develop political strategies.

Two problems typically arise while studying sensitive topics, (1) nonresponse rate increase and (2) social desirability bias (SDB), which is defined as the tendency of answering questions in a socially acceptable fashion rather than answering truthfully, occurs. Nonresponse rates can be reduced by utilizing some strategies such as

using advance letters, offering incentives, using more experienced interviewers, and making the topic salient to respondents [2–6]. However, if these strategies are more effective for some particular subpopulations compared to others, a reduction in nonresponse can in reality increase nonresponse bias. Statistical techniques can also be used to minimize the effects of unit nonresponse after the data is collected [7]; yet, none of these approaches can prevent SDB.

Many researchers have suggested using self-administered modes such as mail, web, computer-assisted self-interviewing (CASI), audio computer-assisted self-interviewing (ACASI), telephone audio computer-assisted self-interviewing (T-ACASI), or touchtone data entry (TDE) in order to reduce SDB [8–11]. However, self-administered modes have their own drawbacks. For example, all self-administered surveys are known to be susceptible to produce low-quality data since they lack the interviewer feedback to help clarify the questions when respondents do not understand them. As another example, it is known that computer-based surveys, such as CASI or ACASI, are mostly completed by younger, more computer savvy respondents, which potentially could introduce bias to estimates. T-ACASI, on the other hand, mainly suffers from high break-offs. There are additional issues about the feasibility of utilizing T-ACASI in survey tools with the elderly [12]. In addition, when surveying disadvantaged populations, self-administered surveys might not be a viable option. In fact, illiteracy, poor vision, respondent preference, or other reasons can cause self-administration not to occur: in a self-administered component of a computer-assisted personal survey, only 79% of CASI cases were actually fully self-administered [13]. For an extensive review of advantages and disadvantages of some of the common survey modes, see Smith and Kim [14].

An effective, alternative way to improve response rates and prevent SDB simultaneously is to increase the perceived privacy of the respondents. If the respondents' privacy can be guaranteed, then their tendency to refuse to participate and/or provide untruthful answers would decrease. All indirect questioning techniques aim to achieve this goal via different approaches. In the next section, we review some of the indirect questioning techniques that have been developed to increase the perceived privacy of the respondents, where the characteristic under study (the outcome) is binary in nature. Some of these techniques explained here have been extended to the cases where the characteristic under study is quantitative or polychotomous in nature; here we will only focus on the binary outcomes, such as in yes/no type of questions. Note that all indirect questioning techniques have produced extensive research areas: they all have been modified and/or extended since they have been first introduced; here we briefly summarize the most important ones and present their main aspects for conciseness.

2. Indirect questioning techniques

Several indirect surveying methodologies have been developed to increase respondents' confidentiality when the characteristic under study has a sensitive nature. Among them, the most commonly used ones for binary outcomes are, namely, unmatched count technique (UCT), network scale-up technique (NST), nonrandomized response technique (NRRT), and randomized response technique (RRT).

UCT, which is also called the item count technique, was first introduced by Raghavarao and Federer [15] with the name "block total response procedure," but it was formally developed by Miller [16] in the form that we use today. Since then, UCT has been applied by many researchers such as Miller et al. [17], LaBrie and Earleywine [18], Biemer and Brown [19], Wolter and Laier [20], Gervais and Najle [21], and so forth. UCT provides privacy by embedding a sensitive behavior (which is of interest) within several nonsensitive behaviors. All nonsensitive behaviors and

sensitive behavior should be binary outcomes (yes/no). In applying the technique, survey participants are randomly divided into two groups. Individuals in one group are provided with a list of nonsensitive behaviors (say, k items), whereas individuals in the other group are provided with the same list of nonsensitive behaviors plus one sensitive behavior that is under study (in total $k + 1$ items). Participants in both groups are asked to report the *total number* of activities that they have engaged in. The prevalence of the sensitive behavior is then estimated by calculating the difference between the two independent group means. Several researchers have modified and/or extended UCT in recent years such as Tsuchiya [22], Chaudhuri and Christofides [23], Hussain et al. [24], Ibrahim [25], and so on. As an example on how to apply UCT, consider the study of Zimmerman and Langer [26] in which the technique was utilized to improve the prevalence estimates of marihuana use and ever having had sex with a person of the same gender among the tenth grade students in the Miami-Dade County Public Schools. In their study, the sample (1524 students) was randomly assigned to receive one of three lists on the survey: 33% of the respondents received a list with four nonsensitive items only; 33% received a list with the four nonsensitive items plus the sensitive item about sex with someone of the same gender; 33% received a list with the four nonsensitive items plus the sensitive item about ever using marihuana. Their UCT estimate of prevalence of having sex with someone of the same gender was 80% higher than the prevalence from direct questioning (DQ) (15.9 vs. 8.7%); similarly, their UCT estimate of marihuana use was 21% higher than the prevalence from DQ (20.6 vs. 16.7%). Assuming that for socially undesirable behaviors, higher prevalence estimates are more accurate than the lower prevalence estimates; Zimmerman and Langer [26] concluded that their results demonstrate the superiority of UCT compared to DQ.

The NST was first proposed by Bernard et al. [27] in order to estimate the size of a population at risk and was first used to get an estimate of the number of victims killed in the 1985 Mexico City earthquake [28]. The method was later refined and used to estimate the HIV-seropositive persons in the USA [29]. NST basically involves two steps: (1) the personal network size of the members of a random sample of a population is estimated and (2) an estimate of the number of members of the hidden subpopulation is obtained using the information from step 1. The method heavily relies on the assumption that people's social networks on average are representative of the general population; for example, if respondents report knowing 500 people on average, five of whom are sex workers, we can estimate that 1% of the general population is sex workers. The estimated prevalence is then combined with known information about the size of the general population, say the population of the USA, to produce an estimate for the number of people in the USA who are sex workers (see Russell et al. [30] for more details on NST and its limitations).

NRRT was first introduced by Swensson [31] and later modified by Takahasi and Sakasegawa [32]. The main idea behind Swensson's NRRT was to combine a nonsensitive behavior with a sensitive behavior in the same question so that it would not be possible for the interviewer to know which behavior is being responded with a "yes" answer; and therefore, respondents are provided with some level of privacy. Swensson's NRRT requires two independent samples to calculate the estimate of the sensitive characteristic; for this purpose, survey participants can be randomly divided into two groups. Let U be the nonsensitive behavior, such as being married, and let A be the sensitive behavior under study, such as abusing prescription drugs. Let us demonstrate the combined question using **Table 1** given below.

Respondents in the first group receive the question "Do you belong to one of the groups a, b, or c in the table?" whereas respondents in the second group receive the question "Do you belong to one of the groups a, b, or d in the table?" Realize that the probability of belonging to groups a, b, c, or d is $P(A = 1, U = 1)$, $P(A = 1, U = 0)$, $P(A = 0, U = 1)$, or $P(A = 0, U = 0)$, respectively. If we denote the unknown

A (Do you use prescription pain relievers without a doctor's prescription?)	U (Are you married?)	
	Yes	No
Yes	a	b
No	c	d

Table 1.
Demonstration of Swensson's NRRT.

proportion of the population members who have the sensitive characteristic A, i.e., the prevalence of the sensitive characteristic, by π_A, and let

$$X_{ij} = \begin{cases} 1, \text{ if the respondent } i \text{ replies with a } ''yes'' \text{ in sample } j \\ 0, \text{ if the respondent } i \text{ replies with a } ''no'' \text{ in sample } j \end{cases}$$

where $0 \leq \pi_A \leq 1$, $i = 1, 2..., n_j$, and $j = 1, 2$, then the probability of getting a "yes" response from sample 1 becomes

$$P(X_{i1} = 1) = P(A = 1, U = 1) + P(A = 1, U = 0) + P(A = 0, U = 1) \quad (1)$$

and the probability of getting a "yes" response from sample 2 becomes

$$P(X_{i2} = 1) = P(A = 1, U = 1) + P(A = 1, U = 0) + P(A = 0, U = 0). \quad (2)$$

Since,

$$P(A = 1, U = 1) + P(A = 1, U = 0) = P(A = 1),$$

and

$$P(A = 0, U = 1) + P(A = 0, U = 0) = P(A = 0),$$

using the Eqs. (1) and (2), we can write

$$P(X_{i1} = 1) + P(X_{i2} = 1) = 2P(A = 1) + P(A = 0),$$

or

$$P(A = 1) = P(X_{i1} = 1) + P(X_{i2} = 1) - 1. \quad (3)$$

From Eq. (3), π_A can be estimated using the proportions of "yes" responses, which are calculated from the samples as:

$$\hat{\pi}_{AS} = \hat{\pi}_1 + \hat{\pi}_2 - 1 \quad (4)$$

where $\hat{\pi}_j = \sum_{i=1}^{n_j} X_{ij}/n_j, j = 1, 2$. One can easily show that the estimator given by (4) is an unbiased estimator of π_A:

$$E(\hat{\pi}_{AS}) = \frac{1}{n_1} E\left(\sum_{i=1}^{n_1} X_{i1}\right) + \frac{1}{n_2} E\left(\sum_{i=1}^{n_2} X_{i2}\right) - 1$$

$$= \frac{1}{n_1} \sum_{i=1}^{n_1} P(X_{i1} = 1) + \frac{1}{n_2} \sum_{i=1}^{n_2} P(X_{i2} = 1) - 1$$

$$= 2P(A = 1) + P(A = 0) - 1$$

$$= \pi_A.$$

The variance of the estimator given in Eq. (4) is derived in a few steps:

$$Var(\hat{\pi}_{AS}) = \frac{1}{n_1^2} Var\left(\sum_{i=1}^{n_1} X_{i1}\right) + \frac{1}{n_2^2} Var\left(\sum_{i=1}^{n_2} X_{i2}\right)$$

$$= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \{P(X_{i1}=1)[1-P(X_{i1}=1)]\} + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \{P(X_{i2}=1)[1-P(X_{i2}=1)]\}$$

$$= \frac{P(X_{i1}=1)[1-P(X_{i1}=1)]}{n_1} + \frac{P(X_{i2}=1)[1-P(X_{i2}=1)]}{n_2},$$

which can be estimated by

$$\widehat{Var}(\hat{\pi}_{AS}) = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}. \tag{5}$$

Swensson's NRRT was later modified by Takahasi and Sakasegawa [32] as follows: In the first stage, all respondents are asked a nonsensitive binary question, such as "If you have to choose between adopting a cat or a dog, which would you prefer?" but directed not to report their answers to the interviewer (replies are silent). In the second stage, the sensitive behavior is combined with the previous nonsensitive question and asked the respondents in the format given below:

- If you are a dog person, and use prescription pain relievers without a doctor's prescription, say 0.

- If you are a dog person, and do not use prescription pain relievers without a doctor's prescription, say 1.

- If you are a cat person, and use prescription pain relievers without a doctor's prescription, say 1.

- If you are a cat person, and do not use prescription pain relievers without a doctor's prescription, say 0.

In this NRRT, to be able to obtain the estimates, the nonsensitive and sensitive behaviors need to be independent (U and A should not be related). As we explain in the next section, if the nonsensitive binary question's prevalence in the population is known, then Takahasi and Sakasegawa's NRRT is equivalent to Warner's model. If the nonsensitive binary question's prevalence in the population is not known, then this method requires the use of two independent samples. By relaxing the assumption of independence of the nonsensitive and sensitive behaviors, Takahasi and Sakasegawa [32] proposed an additional NRRT as well, where the nonsensitive behavior has three outcomes instead of two; as a result, in this NRRT, three independent samples are required instead of two independent samples. For a comprehensive review of NRRTs, interested readers are referred to Tian and Tang [33].

Different than the previous indirect questioning techniques, RRTs provide confidentiality by utilizing a *randomization device*, such as a deck of playing cards, a pair of dice, or a spinning game wheel while asking the sensitive question. The interviewer does not observe which playing card is chosen or what numbers the dice show, thus a respondent's privacy is absolutely guaranteed by the RRT process because the interviewer cannot know which statement the participant is responding to. In the next section, we review some of the main RRTs that have been proposed as well as some of the recently developed techniques.

3. Randomized response techniques

The first RRT was introduced by Warner [34]. Warner's RRT asks the sensitive question by providing respondents a randomization device with two statements on it that appear with known probabilities θ and $1-\theta$. For example, consider a spinning game wheel which divided into two sectors with areas θ and $1-\theta$. The first statement on the device possesses the sensitive characteristic, and the other statement is simply the complementary of the first one. Let us suppose that the sensitive characteristic is being HIV+. The two statements on the wheel would read "I am HIV+" and "I am HIV-" (see **Figure 1**). The respondents spin the wheel in private and answer truthfully with a "yes" or "no" to the statement on which the arrowhead lands. Since the spinning process is not observed by the interviewer, this model assures privacy protection for the respondents [34].

Let the unknown proportion of the population members who have the sensitive characteristic A be denoted by π_A, and let

$$X_i = \begin{cases} 1, \text{ if the respondent } i \text{ replies with a }"yes" \\ 0, \text{ if the respondent } i \text{ replies with a }"no" \end{cases} \tag{6}$$

where $0 \leq \pi_A \leq 1$ and $i = 1, 2..., n;$ then the probability of getting a "yes" response can be written as

$$P(X_i = 1) = \pi_A \theta + (1 - \pi_A)(1 - \theta), \tag{7}$$

and, from Eq. (7), an unbiased estimator of π_A can be derived as

$$\hat{\pi}_{AW} = \frac{\hat{\pi} + \theta - 1}{2\theta - 1},$$

where $\theta \neq 0.5$ and $\hat{\pi} = \sum_{i=1}^{n} X_i / n$. The variance of Warner's estimator can be derived as [34]

$$Var(\hat{\pi}_{AW}) = \frac{\pi_A(1 - \pi_A)}{n} + \frac{\theta(1 - \theta)}{n(2\theta - 1)^2}. \tag{8}$$

Now, let us compare Swensson's NRRT with Warner's model. For simplicity, let us assume that $n_1 = n_2 = n/2$ and the prevalence of U (say, the proportion of the married people in the population) is known from previous studies. Let us also assume that U and A are not related behaviors. In Swensson's NRRT, if we denote the prevalence of U by p, the probability of getting a "yes" response from samples 1 and 2 can be written as
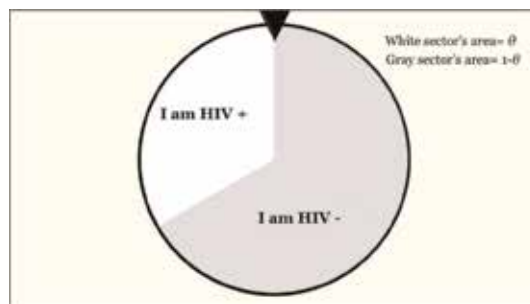
Figure 1.
Demonstration of the randomization device in the Warner's model.

$$P(X_{i1} = 1) = \pi_A + P(A = 0)P(U = 1)$$
$$= \pi_A + (1 - \pi_A)p$$

and

$$P(X_{i1} = 2) = \pi_A + P(A = 0)P(U = 0)$$
$$= \pi_A + (1 - \pi_A)(1 - p),$$

respectively, assuming that the nonsensitive and sensitive behaviors are independent. Since we have

$$P(X_{i1} = 1)[1 - P(X_{i1} = 1)] = \pi_A(1 - \pi_A)(1 - 2p) + p(1 - \pi_A)[1 - p(1 - \pi_A)]$$

and

$$P(X_{i2} = 1)[1 - P(X_{i2} = 1)] = p(1 - \pi_A)[1 - p(1 - \pi_A)],$$

the variance of $\hat{\pi}_{AS}$ can be simplified as

$$Var(\hat{\pi}_{AS}) = \frac{2}{n}\{P(X_{i1} = 1)[1 - P(X_{i1} = 1)] + P(X_{i2} = 1)[1 - P(X_{i2} = 1)]\}$$
$$= \frac{2}{n}\left\{\pi_A(1 - \pi_A)\left[p^2 + (1 - p)^2\right] + 2(1 - \pi_A)p(1 - p)\right\}. \qquad (9)$$

In order to compare Warner's RRT with Swensson's NRRT, we calculated theoretical relative efficiencies (REs)

$$RE_{\theta,p} = \frac{Var(\hat{\pi}_{AW})}{Var(\hat{\pi}_{AS})}$$

from Eqs. (8) and (9) for various combinations of π_A and $\theta = p$ values using a FORTRAN code and providing our results in **Table 2**. The FORTRAN code used is given in the Appendix.

One can conclude from **Table 2** that Warner's model is more efficient, i.e., has smaller variance, than Swensson's NRRT for small $\theta = p$ values. When $\theta = p \geq 0.25$,

π_A	$\theta = p$								
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.1	0.445	0.489	0.590	0.765	1.067	1.630	2.854	6.360	25.298
0.2	0.495	0.546	0.650	0.828	1.138	1.717	2.978	6.592	26.122
0.3	0.524	0.588	0.702	0.892	1.219	1.831	3.162	6.974	27.577
0.4	0.545	0.624	0.754	0.963	1.320	1.984	3.427	7.558	29.878
0.5	0.564	0.662	0.813	1.052	1.455	2.201	3.818	8.446	33.445
0.6	0.585	0.708	0.891	1.175	1.650	2.527	4.424	9.848	39.145
0.7	0.615	0.775	1.009	1.370	1.969	3.072	5.455	12.263	49.028
0.8	0.667	0.899	1.234	1.749	2.600	4.164	7.541	17.188	69.271
0.9	0.817	1.256	1.892	2.867	4.480	7.444	13.843	32.120	130.806

Table 2.
Relative efficiencies of the estimators from Warner's model and Swensson's NRRT for various π_A and $\theta = p$ values. The shaded cells show the cases where Warner's model is more efficient than Swensson's model.

however, Warner's estimator $\hat{\pi}_{AW}$ loses its efficiency. The reason for this efficiency loss is due to the fact that when $\theta \to 0.5$, the variance in Eq. (8) $\to \infty$. However, keep in mind that small θ values are not preferable in Warner's model since they do not provide sufficient privacy protection [33, 35]. In order to further investigate how Warner's RRT compete with Swensson's NRRT, we also assumed the case where $\theta \neq p$ and calculated the theoretical REs as given above, for the case where π_A values were set to 0.1, 0.2, and 0.3. The results are provided in **Table 3**. Realize that we only present the RE values for $0.05 \leq p \leq 0.5$ and $0.05 \leq \theta \leq 0.45$ since

$$RE_{\theta,p} = RE_{1-\theta,p} = RE_{\theta,1-p} = RE_{1-\theta,1-p}.$$

The FORTRAN code used for **Table 3** can be obtained from the author upon request.

One can observe from **Table 3** that Warner's model is more efficient than Swensson's NRRT only for small θ values. One can also observe that REs become higher when π_A values increase, except for some special cases where $p \leq 0.25$ and $\theta \geq 0.10$. Note that in both **Tables 2** and **3**, REs$\to \infty$ when $\theta = 0.5$.

When the nonsensitive binary question's prevalence in the population is known (say, the prevalence of being a dog person is $1 - \theta$, and the prevalence of being a cat person is θ), Takahasi and Sakasegawa's NRRT becomes equivalent to Warner's model. To see this, let the unknown proportion of the population members who have the sensitive characteristic A be denoted by π_A, where $0 \leq \pi_A \leq 1$, and let X_i ($i = 1, 2..., n$) be defined as in Eq. (6). If θ is known, the probability of getting a "1" response from Takahasi and Sakasegawa's NRRT can be written as

$$P(X_i = 1) = (1 - \pi_A)(1 - \theta) + \pi_A\theta, \tag{10}$$

which is equal to the probability of getting a "yes" response from Warner's model, i.e., Eq. (7) and thus, provides the same estimator when solved for π_A.

There have been many different RRTs developed since Warner's original method, such as in Greenberg et al. [36, 37], Gupta [38], Gupta et al. [39–41], Yu et al. [42], Sihm et al. [43], Gupta and Shabbir [44], and so on. Efforts specifically have been made to improve the efficiency of the technique by reducing the variance and thus the confidence intervals, because, the primary disadvantage of the Warner's model, and in fact of all RRTs, is that the variances of the estimators are higher than the ones that could be obtained from DQ [35, 45]; for a comprehensive review of RRTs, interested readers are referred to Chaudhuri and Mukerjee [35], Chaudhuri [46], and Chaudhuri and Christofides [47]. In the next subsection, we will review some real-life applications of RRTs and their comparisons with other surveying techniques.

3.1 Applications of RRTs

Two meta-analyses of 6 validation and 32 comparative studies which utilized RRTs showed that in various settings, the RRT results are superior to those from DQ and become more valid as the sensitivity of a topic increases [48].

The benefits of the RRTs have also been demonstrated by many statistical methodology researchers via theorems and simulation studies; however, their use in large or national surveys has been somewhat limited. In fact, to our knowledge, the only study which applied an RRT on a national level was done by Kirtadze et al. [49] in the country of Georgia. Kirtadze et al. [49] used a multistage cluster sampling and surveyed 4805 respondents to assess under-reporting of drug abuse in the Republic of Georgia. They utilized the unrelated question RRT to ask questions such

p	θ								
	0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**
$\pi_A = 0.1$									
0.05	0.445	0.691	1.049	1.601	2.516	4.200	7.840	18.239	74.394
0.10	0.315	0.489	0.743	1.133	1.781	2.974	5.551	12.913	52.672
0.15	0.251	0.389	0.590	0.901	1.416	2.365	4.414	10.268	41.882
0.20	0.213	0.330	0.501	0.765	1.203	2.008	3.748	8.720	35.567
0.25	0.189	0.293	0.445	0.679	1.067	1.781	3.324	7.733	31.543
0.30	0.173	0.268	0.407	0.621	0.976	1.630	3.043	7.078	28.870
0.35	0.162	0.251	0.382	0.583	0.916	1.529	2.854	6.640	27.085
0.40	0.155	0.241	0.366	0.558	0.877	1.465	2.734	6.360	25.940
0.45	0.151	0.235	0.357	0.544	0.855	1.428	2.666	6.202	25.298
0.50	0.150	0.233	0.354	0.540	0.848	1.417	2.644	6.152	25.091
$\pi_A = 0.2$									
0.05	0.495	0.681	0.952	1.369	2.061	3.334	6.086	13.949	56.409
0.10	0.397	0.546	0.763	1.098	1.653	2.675	4.883	11.192	45.258
0.15	0.338	0.465	0.650	0.935	1.408	2.278	4.158	9.530	38.537
0.20	0.300	0.412	0.576	0.828	1.247	2.018	3.684	8.443	34.142
0.25	0.273	0.376	0.525	0.756	1.138	1.841	3.360	7.700	31.138
0.30	0.255	0.351	0.490	0.705	1.061	1.717	3.134	7.183	29.046
0.35	0.242	0.333	0.466	0.670	1.008	1.632	2.978	6.826	27.604
0.40	0.234	0.322	0.450	0.647	0.974	1.576	2.876	6.592	26.659
0.45	0.229	0.315	0.441	0.634	0.954	1.544	2.819	6.460	26.122
0.50	0.228	0.313	0.438	0.630	0.948	1.534	2.800	6.417	25.948
$\pi_A = 0.3$									
0.05	0.524	0.683	0.916	1.275	1.871	2.967	5.336	12.103	48.646
0.10	0.450	0.588	0.788	1.097	1.610	2.553	4.591	10.412	41.851
0.15	0.401	0.523	0.702	0.977	1.433	2.273	4.087	9.270	37.259
0.20	0.366	0.478	0.641	0.892	1.309	2.075	3.732	8.465	34.024
0.25	0.341	0.445	0.597	0.831	1.219	1.933	3.477	7.886	31.695
0.30	0.323	0.422	0.565	0.787	1.154	1.831	3.292	7.468	30.014
0.35	0.310	0.405	0.543	0.756	1.109	1.758	3.162	7.172	28.826
0.40	0.302	0.394	0.528	0.735	1.078	1.710	3.075	6.974	28.032
0.45	0.297	0.387	0.520	0.723	1.061	1.682	3.025	6.861	27.577
0.50	0.295	0.385	0.517	0.719	1.055	1.673	3.009	6.824	27.429

Table 3.
Relative efficiencies of the estimators from Warner's model and Swensson's NRRT for various θ and p values when $\pi_A = 0.1, 0.2, 0.3$. The shaded cells show the cases where Warner's model is more efficient than Swensson's model.

as "During the last 12 months, have you taken hashish or marihuana?" They found that all RRT estimates for prevalence of controlled substance use were higher than the DQ estimates, which indicates under-reporting with DQ. For example, lifetime cannabis use estimate was 88.24% higher from RRT than from DQ. Kirtadze et al. [49],

however, did not use a gold standard such as urinalysis and thus did not know the "true" value of prevalence of illegal drug use in the Republic of Georgia's study population.

Although not nationwide, there have been other researchers who incorporated RRTs in their surveys. Fisher et al. [50], for example, used forced RRT to estimate the prevalence of substance use and sexual activity of high school students who enrolled to their clinic. They compared their results with the ones from a non-anonymous questionnaire which was completed by the same students earlier the same academic year. While RRT provided higher rates for substance use-related questions with respect to the DQ, it provided similar rates for sexual activity-related questions (36% from DQ vs. 31% from RRT). Fisher et al. [50] concluded that admitting to sexual activity in the school setting might carry less stigma and perceived risk than admitting to marihuana or cocaine use [50]. We suggest, however, that this result might indicate that the high school students who participated in the study overestimated their sexual behavior when asked directly to live up to peer acceptance; in other words, we suggest that forced RRT corrected the over-reporting.

In another study, Srivastava et al. [51] used Warner's RRT to assess the extent of sexual abuse among children in several districts of Uttar Pradesh state of India. They found that the estimates from RRT were higher than the national estimates obtained from Ministry of Women and Child Development, Government of India, which is an indicator of potential under-reporting with DQ.

In a more recent study, Chhabra et al. [52] used convenience sampling and asked the question "Have you ever been a victim of sexual abuse by a friend or family member?" to 585 students in a college in Delhi, India. They divided their sample into three equal randomly selected groups and asked the sexual abuse-related question using (1) DQ, (2) the RRT proposed by Sihm et al. [43], and (3) the confidential method, which was their gold standard to compare their results. The prevalence of sexual abuse was 14% with the gold standard, 8% with the DQ method, and 12% with the RRT. Note that their confidential method, however, was also a surveying technique in which the participants wrote down their answers and put them into a closed box.

3.2 Inflated variance

Although all RRTs lead to unbiased (i.e., accurate) estimates of the sensitive characteristic of interest, their variances are larger than the ones from the DQ technique. Thus, the price for using an RRT instead of the DQ is the inflated variance, which is due to the randomization process. Consequently, if the question of interest is not considered to be really sensitive by most of the respondents in a specific population, using an RRT instead of the DQ inflates the variance of the estimates unnecessarily. Besides, it is known that there are cultural or social differences in the extent to which topics are perceived as sensitive. For example, smoking marihuana is a less threatening topic in the Netherlands than in the USA, or questions regarding education are considered to be sensitive in Sweden [1]. Similarly, questions regarding HIV status or sexual practices might not be considered to be sensitive by patients who visit an HIV clinic for treatment. Unfortunately, once an RRT is incorporated within a survey, even if the question of interest turns out to be not sensitive *after* collecting the data, one has to proceed with the inefficient estimate obtained from the RRT. Motivated by the fact that *if researchers do not have a priori knowledge about the sensitivity level of a question in a specific population, then they should be able to select between the estimators from an RRT and the DQ*; Ardah and Oral [45] proposed a novel two-stage RRT for a binary response where they utilized

Warner's RRT in the second stage. Their model provides unbiased estimators of both prevalence of the sensitive characteristic and the proportion of cheating in the population simultaneously and allows one to obtain better estimates by avoiding the unnecessary penalty, i.e., the inflated variance, if the question of interest turns out to be significantly not sensitive. Although there are some other RRTs which can estimate the sensitivity level of a question, such as in Gupta [38], none of these RRTs have the ability to avoid the inflated variance if the question's sensitivity level turns out to be low.

Ardah and Oral [45] denoted the unknown true proportion of population members that have the sensitive characteristic A by π_A, where $0 \leq \pi_A \leq 1$. They showed that, in a self-protective no-saying model (i.e., when a respondent cheats, s/he always answers in favor of the least stigmatizing category (see [53])) if one uses the DQ, the proportion of respondents who have the sensitive characteristic A becomes

$$\pi_A = \pi_D + \pi_C, \tag{11}$$

where π_D is the proportion of respondents who reply with a "yes" to the sensitive question when asked directly and π_C is the proportion of respondents who cheat (i.e., the proportion of respondents who reply with a "no" to the sensitive question even though they have the characteristic A). Realize that in Eq. (11), π_C is the SDB, and if all respondents answer truthfully in the DQ, i.e., if $\pi_C = 0$, then $\pi_A = \pi_D$. Unfortunately, the SDB, i.e., the proportion of cheaters, cannot be measured using the DQ. Ardah and Oral's [45] proposed model, however, allows one to both estimate it and select the unbiased estimator with the minimum variance depending on the sensitivity level of the question (or proportion of cheating in the population). Ardah and Oral [45] utilized unrelated question RRT in the second stage of their model as well; in fact, their model can easily be extended to any desired RRT [45].

4. Conclusion

Numerous indirect questioning techniques have been developed to ask survey questions on sensitive topics or stigmatizing characteristics. All indirect questioning techniques have their own limitations: In applying the UCT, only one of the group members provide the information on the sensitive characteristic of interest [25]. NST is known to suffer from recall bias, barrier effects, transmission error, and response bias [54]. NRRTs can be vulnerable to cheating due to distrust as much as the RRTs [55]. RRTs have some disadvantages as well: integrating a randomization device into a survey tool might not be practical in some situations, such as when researchers plan to use venue sampling to reach LGBT community members in gay bars or clubs. RRTs are also known not to work well if the respondents do not understand the process and/or do not follow the instructions properly. Besides, there might be cultural, social, or personal differences in the extent to which topics are perceived to be sensitive. Thus, we suggest that researchers should select the optimal surveying technique by considering various aspects of their study, such as the target population, sensitivity level of the question, available resources, and practicality of integrating a specific technique, at once. As in Erdmann [55], we also suggest that researchers should not rely on the more-is-better assumption, which is assuming that the higher prevalence estimates are more accurate than the lower prevalence estimates, in comparing different techniques; instead, we suggest to use a valid gold standard (such as urinalysis or a lie detector) for comparisons, whenever possible, perhaps using a small subsample.

Appendix

```
!**************************************************************************
! This code was written to calculate the relative efficiencies in Table 2
!**************************************************************************
INCLUDE 'link_fnl_static.h'
REAL(8) THETA,C, PIA, K,E, VAR_RRT, VAR_NRRT
REAL(8) R(21,21)
INTEGER I,J,ITENO,N
OPEN (3,FILE="C:\Users\Oral\Documents\Fortran results\Table2.txt")
WRITE(3,*) ""
WRITE(3,*) ""
WRITE(3,50) "Calculated Relative Efficiencies for various p=theta and Popula-
tion Proportion (Pi_A) combinations"
WRITE(3,*)""
WRITE(3,*)"*****RE MATRIX*****"
WRITE(3,*)""
WRITE(3,*)"p=Theta value "
DO K=0.0,1.05,0.05
WRITE(3, 100,advance='no') K
END DO
WRITE(3, *) "
DO J=1,148
WRITE(3, 200,advance='no') '_'
END DO
WRITE(3, *) "
THETA=0.0
PIA=0.0
DO I=1,11
THETA=0.0
DO J=1,21
WRITE (*,*) 'theta=p=', THETA, 'Pi_A=', PIA
VAR_RRT=(PIA*(1-PIA))+((THETA*(1-THETA))/(((2*THETA)-1)**2.0))
VAR_NRRT=2*(((((PIA*(1-PIA)))*(1-(2*THETA)+(2*(THETA**2.0))))))+(2*(1-
PIA)*THETA*(1-THETA)))
WRITE (*,*) 'VAR_RRT=',VAR_RRT, 'VAR_NRRT=', VAR_NRRT
R(I,J)=VAR_RRT/VAR_NRRT
THETA=THETA+0.05
WRITE(3,100,advance='no') R(I,J);
END DO
WRITE (*,*)"
WRITE(3, *) "
PIA=PIA+0.1
END DO
100 FORMAT (F15.5,1X)
200 FORMAT (A1)
CLOSE(3)
END
```

Author details

Evrim Oral
LSUHSC School of Public Health, Biostatistics program, New Orleans, LA, USA

*Address all correspondence to: eoral@lsuhsc.edu

IntechOpen

References

[1] Lensvelt-Mulders G. Surveying sensitive topics. In: de Leeuw ED, Hox JJ, Dillman DA, editors. International Handbook of Survey Methodology. New York: LEA, Taylor & Francis; 2008. pp. 1-17

[2] DeLeeuw E, Callegaro M, Hox J, Korendijk E, Lensvelt-Mulders G. The influence of advanced letters on response in telephone surveys: A meta-analysis. Public Opinion Quarterly. 2007;**71**(3):413-443

[3] Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey Methodology. 2nd ed. Hoboken, NJ: Wiley; 2009

[4] Link M, Mokdad A, Town M, Weiner J, Roe D. Improving Response Rates for the BRFSS: Use of Lead Letters and Answering Machine Messages. Paper presented at the annual conference of the American Association for Public Opinion Research, Nashville, TN; 2003

[5] Singer E, Groves RM, Dillman DA, Eltinger JL, Little RJA, editors. The Use of Incentives to Reduce Nonresponse in Household Surveys. Wiley-Interscience; 2002. pp. 163-177

[6] Spiers S, Oral E, Fontham E, Peters ES, Mohler JL, Bensen JT, et al. Modelling attrition and nonparticipation in a longitudinal study of prostate cancer. BMC Medical Research Methodology. 2018;**18**:60. DOI: 10.1186/s 12874-018-0518-6

[7] Oral E, Simonsen N, Brennan C, Berken J, Su LJ, Mohler JL, et al. Unit nonresponse in a population-based study of prostate cancer. PLoS One. 2016;**11**(12):e0168364. DOI: 10.1371/journal.pone.0168364

[8] Lessler JT, O'Reilly JM. Mode of interview and reporting of sensitive issues: Design and implementation of audio computer assisted self interviewing. NIDA Research Monograph. 1997;**167**:366-382

[9] van Griensven F, Naorat S, Kilmarx PH, et al. Palmtop-assisted self-interviewing for the collection of sensitive behavioral data: Randomized trial with drug use urine testing. American Journal of Epidemiology. 2006;**163**(3):271-278

[10] Lind LH, Schober MF, Conrad FG, Reichert H. Why do survey respondents disclose more when computers ask the questions? Public Opinion Quarterly. 2013;**77**:888-935

[11] Schober MF, Conrad FG, Antoun C, Ehlen P, Fail S, Hupp AL, et al. Precision and disclosure in text and voice interviews on smartphones. PLoS One. 2015;**10**(6):e0128337

[12] Beach SR, Schulz R, Degenholtz HB, Castle NG, Rosen J, Fox AR, et al. Using audio computer-assisted self-interviewing and interactive voice response to measure elder mistreatment in older adults: Feasibility and effects on prevalence estimates. Journal of Official Statistics. 2010;**26**(3):507-533

[13] Couper MP, Rowe B. Evaluation of a computer-assisted self-interview component in a computer-assisted personal interview survey. Public Opinion Quarterly. 1996;**60**:89-105

[14] Smith TW, Kim J. A review of survey data collection modes: With a focus on computerizations. Sociological Theory and Methods. 2015;**20**(2):185-200

[15] Raghavarao D, Federer WT. Block total response as an alternative to the randomized response method in surveys. Journal of the Royal Statistical Society: Series B: Methodological. 1979;**41**:40-45

[16] Miller JD. A new survey technique for studying deviant behavior [PhD thesis]. The George Washington University; 1984

[17] Miller J, Cisin I, Harrell A. A new technique for surveying deviant behavior: Item-count estimates of marijuana, cocaine, and heroin. Paper presented at the Annual Meeting of the American Association for Public Opinion Research. St. Petersburg, FL; 1986

[18] LaBrie JW, Earleywine M. Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. The Journal of Sex Research. 2000;**37**(4):321-326

[19] Biemer P, Brown G. Model-based estimation of drug use prevalence using item count data. Journal of Official Statistics. 2005;**21**(2):287-308

[20] Wolter F, Laier B. The effectiveness of the item count technique in eliciting valid answers. To sensitive questions: An evaluation in the context of self-reported delinquency. Survey Research Methods. 2014;**8**(3):153-168

[21] Gervais WM, Najle MB. How many atheists are there? Social Psychological and Personality Science. 2018;**9**(1):3-10. DOI: 10.1177/1948550617707015

[22] Tsuchiya T. Domain estimators for the item count technique. Survey Methodology. 2005;**31**(1):41-51

[23] Chaudhuri A, Christofides TC. Item count technique in estimating the proportion of people with a sensitive feature. Journal of Statistical Planning and Inference. 2007;**137**:589-593

[24] Hussain Z, Ali Shah E, Shabir J. An alternative item count technique in sensitive surveys. Revista Colombiana Estadística. 2012;**35**:39-54

[25] Ibrahim F. An alternative modified item count technique in sampling survey. International Journal of Statistics and Applications. 2016;**6**: 177-187

[26] Zimmerman RS, Langer LM. Improving estimates of prevalence rates of sensitive behaviors: The randomized lists technique and consideration of self-reported honesty. The Journal of Sex Research. 1995;**32**(2):107-117

[27] Bernard HR, Johnsen EC, Killworth PD, Robinson S. Estimating the size of an average personal network and of an event subpopulation. In: Kochen M, editor. The Small World. Norwood, NJ: Albex Pub. Corp.; 1989. pp. 159-175

[28] Bernard HR, Johnsen EC, Killworth PD. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. Social Science Research. 1991;**20**: 109-121

[29] Killworth PD, Johnsen EC, McCarty C, Shelley GA, Bernard HR. A social network approach to estimating seroprevalence in the United States. Social Networks. 1998;**20**:23-50

[30] Russell HB, Hallett T, Iovita A, Johnsen EC, Lyerla R, McCarty C, et al. Counting hard-to-count populations: The network scale-up method for public health sexually transmitted infections. 2010;**86**(Supp. 2):ii11-ii15

[31] Swensson B. Combined questions: A new survey technique for eliminating evasive answer bias (I)—Basic theory. Report No. 70 of the Errors in Surveys Research Project. Institute of Statistics, University of Stockholm; 1974

[32] Takahasi K, Sakasegawa H. A randomized response technique without making use of any randomizing device. Annals of the Institute of Statistical Mathematics. 1977;**29**(1):1-8

[33] Tian GL, Tang M-L. Incomplete Categorical Data Design:

Non-Randomized Response Techniques for Sensitive Questions in Surveys. Boca Raton, FL: Chapman & Hall/CRC; 2014

[34] Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association. 1965; **60**:63-69

[35] Chaudhuri A, Mukerjee R. Randomized Response: Theory and Techniques. Statistics: Textbooks and Monographs. Vol. 85. New York: Marcel Dekker, Inc.; 1988

[36] Greenberg RG, Abul-Ela ALA, Simmons WR, Horvitz DG. The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association. 1969;**64**(326):520-539

[37] Greenberg RG, Keubler RT, Abernathy JR, Horvitz DG. Application of randomized response technique in obtaining quantitative data. Journal of the American Statistical Association. 1971;**66**:243-250

[38] Gupta SN. Qualifying the sensitivity level of binary response personal interview survey questions. Journal of Combinatorics, Information & System Sciences. 2001;**26**:101-109

[39] Gupta SN, Gupta RC, Singh S. Estimation of sensitivity level of personal interview survey questions. Journal of Statistical Planning and Inference. 2002;**100**:239-247

[40] Gupta SN, Thornton B, Shabbir J, Singhal S. A comparison of multiplicative and additive optional RRT models. Journal of Statistical Theory and Applications. 2006;**5**: 226-239

[41] Gupta SN, Shabbir J, Sehra S. On the estimation of population mean and sensitivity in a two-stage optional randomized response model. Journal of

the Indian Society of Agricultural Statistics. 2010;**61**:164-168

[42] Yu B, Jin Z, Tian J, Gao G. Estimation of sensitive proportion by randomized response data in successive sampling. Computational and Mathematical Methods in Medicine. 2015;**18**(2015):172918

[43] Sihm JS, Chhabra A, Gupta S. An optional unrelated question RRT model. Involve: A Journal of Mathematics. 2016;**2**(9):195-209

[44] Gupta SN, Shabbir J. Sensitivity estimation for personal interview survey questions. Statistica. 2004;**64**:643-653

[45] Ardah IH, Oral E. Model selection in randomized response techniques for binary responses. Communication in Statistics-Theory and Methods. 2018; **47**(14):3305-3323

[46] Chaudhuri A. Randomized Response and Indirect Questioning Techniques in Surveys. Boca Raton, FL: Chapman and Hall/CRC Taylor and Francis Group. 2011

[47] Chaudhuri A, Christofides TC. Indirect Questioning in Sample Surveys. Berlin, Heidelberg: Springer-Verlag; 2013. DOI: https://doi.org/10.1007/987-3-642-36276-7

[48] Lensvelt-Mulders G, Hox JJ, van der Heijden P, Maas C. Meta-analysis of randomized response: 35 years of validation studies. Sociological Methods & Research. 2005;**33**:319-348

[49] Kirtadze I, Otiashvili D, Tabatadze M, Vardanashvili I, Stutua L, Zabransky T, et al. Republic of Georgia estimates for prevalence of drug use: Randomized response techniques suggest under-estimation. Drug and Alcohol Dependence. 2018;**187**:300-304

[50] Fisher M, Kupferman LB, Lesser M. Substance use in a school-based clinic

population use of the randomized response technique to estimate prevalence. The Journal of Adolescent Health. 1992;**13**:281-285

[51] Srivastava R, Nigam AK, Singh N. Application of randomized response techniques in estimation of prevalence of child sexual abuse. Statistics and Applications. 2015;**13**:37-45

[52] Chhabra A, Dass BK, Gupta S. Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question RRT model. The North Carolina Journal of Mathematics and Statistics. 2016;**2**:1-9

[53] van den Hout A, Bockenholt U, Van der Heijden PGM. Estimating the prevalence of sensitive behavior and cheating with a dual design for direct questioning and randomized response. Applied Statistics. 2010;**59**:723-736

[54] Jing L, Lu Q, Cui Y, Yu H, Wang T. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: An exploratory study. Public Health. 2018;**160**:81-86

[55] Erdmann A. Non-randomized response models: An experimental application of the triangular model as an indirect questioning method for sensitive topics. Methods, Data Analyses. 2018;**13**(1):139-167. DOI: 10.12758/mda.2018.07

Chapter 3

Development of Estimation Procedure of Population Mean in Two-Phase Stratified Sampling

Partha Parichha, Kajla Basu and Arnab Bandyopadhyay

Abstract

This article describes the problem of estimation of finite population mean in two-phase stratified random sampling. Using information on two auxiliary variables, a class of product to regression chain type estimators has been proposed and its characteristic is discussed. The unbiased version of the proposed class of estimators has been constructed and the optimality condition for the proposed class of estimators is derived. The efficacy of the proposed methodology has been justified through empirical investigations carried over the data set of natural population as well as the data set of artificially generated population. The survey statistician may be suggested to use it.

1. Introduction

In this present paper we have made use of Auxiliary information extracted from the variables having correlation with study variable. Auxiliary information may be utilized at planning, design and estimation stages to develop improved estimation procedures in sample surveys. Sometimes, information on auxiliary variable may be readily available for all the units of population; for example, tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation and number of beds available in different hospitals may be known well in advance in health care surveys. If such information lacks, it is sometimes, relatively cheap to take a large preliminary sample where auxiliary variable alone is measured, such practice is applicable in two-phase (or double) sampling. Two-phase stratified sampling happens to be a powerful and cost effective (economical) technique for obtaining the reliable estimate in first-phase (preliminary) sample for the unknown parameters of the auxiliary variables. For example, Sukhatme [1] mentioned that in a survey to estimate the production of lime crop based on orchards as sampling units, a comparatively larger sample is drawn to determine the acreage under the crop while the yield rate is determined from a sub sample of the orchards selected for determining acreage.

In order to construct an efficient estimator of the population mean of the auxiliary variable in first-phase (preliminary) sample, Chand [2] introduced a technique

of chaining another auxiliary variable with the first auxiliary variable by using the ratio estimator in the first phase sample. The estimator is known as chain-type ratio estimator. This work was further extended by Kiregyera [3, 4], Tracy et al. [5], Singh and Espejo [6], Gupta and Shabbir [7], Shukla et al. [8], Choudhury and Singh [9], Parichha et al. [10] and among others, where they proposed various chain-type ratio and regression estimators.

In practice, the population may often consist of heterogeneous units. For example, in socio-economic surveys, people may live in rural areas, urban localities, ordinary domestic houses, hostels, hospitals and jail, etc. In such a situation one should carefully study the population according to the characteristics of regions and then apply sampling scheme strata wise independently. This procedure is known as stratified random sampling. It may be noted that most of the developments in two-phase sampling scheme are based on simple random sampling only while limited number of attempts are taken to address the problems of two-phase sampling scheme in the platform of stratified random sampling. It may be also noticeable that the most of the research work on two-phase sampling are producing biased estimates. However, biased becomes a serious drawback in sample surveys. A sampling method is called biased if it systematically favors some outcomes over others. It results in a biased sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected. If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling. For example, telephone sampling is common in marketing surveys. A simple random sample may be chosen from the sampling frame consisting of a list of telephone numbers of people in the area being surveyed. This method does involve taking a simple random sample, but it is not a simple random sample of the target population (consumers in the area being surveyed). It will miss people who do not have a phone. It may also miss people who only have a cell phone that has an area code not in the region being surveyed. It will also miss people who do not wish to be surveyed, including those who monitor calls on an answering machine and don't answer those from telephone surveyors. Thus the method systematically excludes certain types of consumers in the area. It is obvious that the inferences from a biased sample are not as trustworthy as conclusions from a truly random sample.

Encouraged with the above work, we have proposed a class of product to regression chain type estimators in stratified sampling using two auxiliary variables under double sampling. The unbiased version of the proposed class of estimators has been obtained which make the estimation strategy more practicable. The dominance of the proposed estimation strategy over the conventional ones has been established through empirical investigations carried over the data set of natural as well as artificially generated population.

2. Sampling structures and notations

Consider a finite population U = {1, 2,..., N} of N identifiable units divided into *L* homogeneous strata with the hth stratum (h = 1, 2,..., L) having N_h. Let *y* and (*x, z*) be the study variable and two auxiliary variables respectively taking values y_{ih} and (x_{ih}, z_{ih}), respectively, for the unit *i* = 1,2,...N_h of the hth stratum. $\overline{Y} = \sum_{h=1}^{L} \overline{Y}_h W_h, \overline{X} = \sum_{h=1}^{L} \overline{X}_h W_h, \overline{Z} = \sum_{h=1}^{L} \overline{Z}_h W_h$ be population means of the study and the auxiliary variables, and $\overline{Y}_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}, \overline{X}_h = \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}, \overline{Z}_h = \sum_{i=1}^{N_h} \frac{z_{hi}}{N_h}$ be the corresponding stratum means. Here $W_h = \frac{N_h}{N}$ is the known stratum weight.

Let $C_{yh} = \frac{S_{yh}}{\overline{Y}_h}$, $C_{xh} = \frac{S_{xh}}{\overline{X}_h}$ and $C_{zh} = \frac{S_{zh}}{\overline{Z}_h}$ be the coefficients of variation where $S_{y_h} = \sqrt{\frac{\sum_{i=1}^{N_h}(y_{hi}-\overline{Y}_h)^2}{N_h-1}}$, $S_{x_h} = \sqrt{\frac{\sum_{i=1}^{N_h}(x_{hi}-\overline{X}_h)^2}{N_h-1}}$, $S_{z_h} = \sqrt{\frac{\sum_{i=1}^{N_h}(z_{hi}-\overline{Z}_h)^2}{N_h-1}}$ are the population standard deviations in the hth stratum.

Let ρ_{yx_h}, ρ_{yz_h} and ρ_{xz_h} be the correlation coefficients between (y, x), (y, z), and (x, z) respectively in the hth stratum. Chand [2] and Kiregyera [3, 4] discussed a situation in simple random sampling when information on x is unknown but another auxiliary variable z is easily available. It is assumed that population mean of one auxiliary variable z is known in advance and the population mean of the other auxiliary variable x is unknown. We seek to estimate through a two-phase stratified sampling design. Using a simple random sample without replacement (SRSWOR) sampling scheme at each phase, we adopt the double sampling scheme as follows.

i. In the first phase, a preliminary large sample of size n'_h is drawn from the hth stratum of size N_h (h = 1, 2,..., L) and information on the auxiliary variables x and z is observed.

ii. In the second phase, a sub-sample of size n_h is drawn from first phase sample n'_h units from the h th stratum of size N_h and information on both the study variable y and the auxiliary variables x and z is taken.

$\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi}$, $\overline{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi}$, $\overline{z}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}z_{hi}$, $\overline{x}'_h = \frac{1}{n'_h}\sum_{i=1}^{n'_h}x_{hi}$, and $\overline{z}'_h = \frac{1}{n'_h}\sum_{i=1}^{n'_h}z_{hi}$ be the corresponding sample means in the hth stratum.

3. Discussion on existing estimation strategies

The usual stratified mean estimator $(\overline{y}_{st})$ for population mean $(\overline{Y})$, is given by

$$\overline{y}_{st} = \sum_{h=1}^{L}w_h\overline{y}_h \tag{1}$$

The mean square error (*MSE*) of $\overline{y}_{st}$, is given by

$$\text{MES}(\overline{y}_{st}) = \sum_{h=1}^{L}w_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)s_{yh}^2 \tag{2}$$

Motivated with the technique adopted by Chand [2], one may frame the chain ratio-product type estimator in stratified sampling structure as

$$\overline{y}_{RP}^{(h)} = \sum_{h=1}^{L}w_h\overline{y}_h\left(\frac{\overline{X}'_h}{\overline{x}_h}\right)\left(\frac{\overline{Z}_h{}'}{\overline{z}_h}\right) \tag{3}$$

The bias and *MSE* respectively of $\overline{y}^{(h)}{}_{RP}$, to first order of approximation, are obtained as

$$\text{Bias}\left(\overline{y}_{RP}^{(h)}\right) \cong \sum_{h=1}^{L}w_h\overline{y}_h\left[\left(\frac{1}{n_h}-\frac{1}{n'_h}\right)A_{1h}+\left(\frac{1}{n'_h}-\frac{1}{N_h}\right)A_{2h}\right] \tag{4}$$

$$\text{MSE}\left(\overline{y}_{RP}^{(h)}\right) = \sum_{h=1}^{L}w_h^2s_{yh}^2\left[\left(\frac{1}{n_h}-\frac{1}{n'_h}\right)A_{3h}+\left(\frac{1}{n'_h}-\frac{1}{N_h}\right)A_{4h}+\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\right] \tag{5}$$

where

$$A_{1h} = C_{xh}^2 - \rho_{yxh}C_{yh}C_{xh} \text{ and } A_{2h} = C_{zh}^2 - \rho_{yzh}C_{yh}C_{zh}$$

$$A_{3h} = \frac{C_{xh}^2}{C_{yh}^2} - 2\rho_{yxh}\frac{C_{xh}}{C_{yh}} \quad \text{and} \quad A_{4h} = \frac{C_{zh}^2}{C_{yh}^2} - 2\rho_{yzh}\frac{C_{zh}}{C_{yh}}$$

Similarly, inspired with the technique adopted by Choudhary and Sing [9], one may frame the two-phase stratified random sampling estimator in stratified sampling as

$$\overline{y}_{cs}^h = \sum_{h=1}^L w_h \overline{y}_h \left[k_h \left(\frac{\overline{x}_h'}{\overline{x}_h}\right)\left(\frac{\overline{Z}_h'}{\overline{z}_h}\right) + (1 - k_h)\left(\frac{\overline{x}_h'}{\overline{x}_h}\right)\left(\frac{\overline{z}_h'}{\overline{Z}_h}\right) \right] \tag{6}$$

where k_h is constant.

Bias $\left(\overline{y}_{cs}^h\right) \cong \sum_{h=1}^L w_h \overline{y}_h A_{5h}$

$$A_{5h} = (1 - 2k_h)C_{yh}\left[\left(\frac{1}{n_h} - \frac{1}{n_h'}\right)\rho_{yxh}C_{xh} + \left(\frac{1}{n_h'} - \frac{1}{N_h}\right)\rho_{yzh}C_{zh}\right] + k_h\left[\left(\frac{1}{n_h} - \frac{1}{N_h}\right)C_{xh}^2 + \left(\frac{1}{n_h'} - \frac{1}{N_h}\right)C_{zh}^2\right] \tag{7}$$

$$\text{And MSE} \quad \left(\overline{y}_{cs}^h\right)_{min} = \sum_{h=1}^L w_h^2 s_{y_h}^2 \times \left[\left(\frac{1}{n_h} - \frac{1}{N_h}\right) - \frac{\left\{\left(\frac{1}{n_h} - \frac{1}{n_h'}\right)\rho_{yx_h}C_{x_h} - \left(\frac{1}{n_h'} - \frac{1}{N_h}\right)\rho_{yz_h}C_{z_h}\right\}^2}{\left(\frac{1}{n_h} - \frac{1}{N_h}\right)C_{x_h}^2 + \left(\frac{1}{n_h'} - \frac{1}{N_h}\right)C_{z_h}^2}\right] \tag{8}$$

4. Formulation of proposed estimation strategy

Motivated with the earlier work, discussed above, we have constructed a class of product to regression chain type estimators as

$$t_p = \sum_{h=1}^L w_h \overline{y}_h \left\{ k_h \frac{\overline{x}_h'}{\overline{x}_h} + (1 - k_h)\frac{\overline{x}_{id_h}'}{\overline{x}_h} \right\} \tag{9}$$

where $k_h(h = 1, 2, ..., L)$ is a real constant which can be suitably determined by minimizing the M. S. E. of the class of estimator t_p and $\overline{x}_{d_h}' = \overline{x}' + b_{xz_h}(n_h')\left(\overline{Z}_h - \overline{z}_h'\right)$; where $b_{xz_h}(n_h')$ is the regression coefficient between the variables x and z at the hth stratum.

5. Bias and mean square errors of the proposed class of estimator t_p

It can be easily noted that the proposed class of estimators t_p defined in Eqs. (8) is chain product and regression type estimator. Therefore, it is biased estimator for population mean $\overline{Y}$. So, we obtain biases and mean square errors under large sample approximations using the following transformations:

$$\overline{y}_h = \overline{Y}_h(1 + e_1), \overline{x}_h = \overline{X}_h(1 + e_2), \overline{x}_h' = \overline{X}_h(1 + e_3), \overline{z}_h' = \overline{Z}_h(1 + e_4),$$

$$s_{xz_h}' = S_{xz_h}(1 + e_5), \quad s_{z_h'}^2 = S_{z_h}^2(1 + e_6)$$

and $E(e_i) = 0$ for (i = 1, 2,..., 6), e_i for (i = 1, 2,..., 6) are relative error term. Under above transformations the class of estimator t_p may be represented as

$$t_p = \sum_{h=1}^{L} w_h \overline{Y}(1+e_1)\left[(1-k_h)\left\{(1+e_3)(1+e_2)^{-1}\right\}\right.$$
$$\left. + k_h\left\{(1+e_3) - \frac{\overline{Z}_h}{\overline{X}_h}\beta_{xz_h}(e_4 + e_4e_5 - e_4e_6)\right\}(1+e_2)^{-1}\right] \tag{10}$$

We have the following expectations of the sample statistics of two-phase stratified sampling as

$$
\left.
\begin{array}{l}
E(e_1^2) = f_1 C_{y_h}^2, \; E(e_2^2) = f_1 C_{x_h}^2, \; E(e_3^2) = f_2 C_{x_h}^2, \; E(e_4^2) = f_2 C_{z_h}^2 \\[4pt]
E(e_1 e_2) = f_1 \rho_{yx_h} C_{y_h} C_{x_h}, \; E(e_1 e_3) = f_2 \rho_{yx_h} C_{y_h} C_{x_h}, \\[4pt]
E(e_2 e_3) = f_2 C_{x_h}^2, \; E(e_2 e_4) = E(e_3 e_4) = f_2 \rho_{xz_h} C_{x_h} C_{z_h}, \\[4pt]
E(e_4 e_5) = f_2 \dfrac{\mu_{102}}{\overline{Z}_h S_{xz_h}}, \; E(e_4 e_6) = f_2 \dfrac{\mu_{003}}{\overline{Z}_h S_{z_h}^2}, \\[8pt]
E(e_2 e_5) = f_2 \dfrac{\mu_{201}}{\overline{X}_h S_{xz_h}}, \; E(e_2 e_6) = f_2 \dfrac{\mu_{102}}{\overline{X}_h S_{z_h}^2}, \\[8pt]
E(e_1 e_4) = f_2 \rho_{yz_h} C_{y_h} C_{z_h}.
\end{array}
\right\} \tag{11}
$$

where

$$f_1 = \frac{1}{n_h} - \frac{1}{N_h}, \; f_3 = \frac{1}{n_h} - \frac{1}{n_h'}, \; f_2 = \frac{1}{n_h'} - \frac{1}{N_h},$$

$$\mu_{pqr} = \frac{1}{N_h}\sum_{i=1}^{N_h}(x_i - \overline{X}_h)^P(y_i - \overline{Y}_h)^q(z_i - \overline{Z}_h)^r; \; (p, q, r \geq 0)$$

Expanding binomially, using results from Eq. (1) and retaining the terms up to first order of sample size, we have derived the expressions of bias B(.) and mean square error M(.) of the class of estimators t_p as

$$B(t_p) = E(t_p - \overline{Y}_h) = \sum_{h=1}^{L} w_h \overline{Y}\left[(1-k_h)b_{xz_h}\frac{\overline{Z}_h}{\overline{X}_h}\left(f_2\frac{S_{xz_h}}{\overline{X}_h\overline{Z}_h} - f_1\frac{S_{yz_h}}{\overline{Y}_h\overline{Z}_h} - f_2\frac{\mu_{102}}{S_{xz_h}\overline{Z}_h} - \frac{\mu_{003}}{S_{z_h}^2\overline{Z}_h}\right)\right.$$
$$\left. + f_3\left(\frac{S_{x_h}^2}{\overline{X}_h^2} - \frac{S_{yx_h}}{\overline{Y}_h\overline{X}_h}\right)\right] \tag{12}$$

$$M(t_p) = E[t_p - \overline{Y}_h]^2 = \sum_{h=1}^{L} w_h \overline{Y}_h^2\left[f_1 C_{y_h}^2 + k_h^2 a + 2k_h b + c\right] \tag{13}$$

where $a = \left(f_2\rho_{xz_h}^2\right)C_{x_h}^2$ and $b = f_2\rho_{yz_h}\rho_{xz_h}C_{y_h}C_{x_h} - \left(f_2\rho_{xz_h}^2\right)C_{x_h}^2$
$c = f_3 C_{x_h}^2 - 2f_3\rho_{yx_h}C_{y_h}C_{x_h} + \left(f_2\rho_{xz_h}^2\right)C_{x_h}^2 - 2f_2\rho_{yz_h}\rho_{xz_h}C_{y_h}C_{x_h}.$

6. Bias reduction for the proposed class of estimators

In recent time serious drawback is bias of an estimator. Therefore, unbiased versions of the proposed classes of estimators are more desirable. Motivated with

this argument and influenced by the bias correction techniques of Tracy et al. [5] and Bandyopadhyay and Singh [11] we proceed to derive the unbiased version of our proposed class of estimator t_p.

From Eq. (12), we observe that the expression of bias of the estimator t_p contains the population parameters such as μ_{003}, μ_{102}, S_{yx_h}, S_{yz_h}, $S_{x_h}^2$, $S_{y_h}^2$, $\overline{Y}_h$, $\overline{X}_h$, S_{yz_h} and $S_{z_h}^2$. Since $S_{z_h}^2$ is known while μ_{003}, μ_{102}, S_{yx_h}, S_{yz_h}, $S_{x_h}^2$, $S_{y_h}^2$, $\overline{Y}_h$, $\overline{X}_h$ and S_{yz_h} are unknown, replacing μ_{003}, μ_{102}, S_{yx_h}, S_{yz_h}, $S_{x_h}^2$, $S_{y_h}^2$, $\overline{Y}_h$, $\overline{X}_h$, by their respective sample estimator (based on the second phase sample of size m) m_{003}, m_{102}, s_{yz_h}, $s_{x_h}^2$, $s_{y_h}^2$, $\overline{y}_h$, $\overline{x}_h$ and s_{yz_h}, we get an estimator of $B(t_p)$ and

$$b(t_p) = \sum_{h=1}^{L} w_h \overline{y}_h \left[(1 - k_h) b_{xz_h} \frac{\overline{z}_h}{\overline{x}_h} \left(f_2 \frac{s_{xz_h}}{\overline{x}_h \overline{z}_h} - f_1 \frac{s_{yz_h}}{\overline{y}_h \overline{z}_h} - f_2 \frac{m_{102}}{s_{xz_h} \overline{z}_h} - \frac{m_{003}}{s_{Z_h}^2 \overline{z}_h} \right) + f_3 \left(\frac{s_{x_h}^2}{\overline{x}_h^2} - \frac{s_{yx_h}}{\overline{y}_h \overline{x}_h} \right) \right].$$

(14)

where $m_{pqr} = \frac{1}{m} \sum_{i=1}^{m} (x_{hi} - \overline{x}_h)^p (y_{hi} - \overline{y}_h)^q (z_{hi} - \overline{z}_h)^r$.

Motivating with the bias reduction techniques of Tracy et al. [5] and Bandyopadhyay and Singh [11], we have derived the unbiased version of the proposed class of estimators t_p to the first order of approximations two-phase stratified sampling.

$$t_p' = t_p - b(t_p)$$

which becomes

$$t_p' = \sum_{h=1}^{L} w_h \left[\overline{y}_h \left\{ k_h \frac{\overline{x}_h'}{\overline{x}_h} + (1 - k_h) \frac{\overline{x}_{id_h}'}{\overline{x}_h} \right\} - \overline{y}_h \left[(1 - k_h) b_{xz_h} \frac{\overline{z}_h}{\overline{x}_h} \left(f_2 \frac{s_{xz_h}}{\overline{x}_h \overline{z}_h} - f_1 \frac{s_{yz_h}}{\overline{y}_h \overline{z}_h} - f_2 \frac{m_{102}}{s_{xz_h} \overline{z}_h} - \frac{m_{003}}{s_{Z_h}^2 \overline{z}_h} \right) \right. \right.$$
$$\left. \left. + f_3 \left(\frac{s_{x_h}^2}{\overline{x}_h^2} - \frac{s_{yx_h}}{\overline{y}_h \overline{x}_h} \right) \right] \right]$$

(15)

Thus, the variance of t_p' to the first order of approximation are obtained as

$$V\left(t_p'\right) = M(t_p) = \sum_{h=1}^{L} \overline{Y}_h^2 \left[f_1 C_{y_h}^2 + k_h^2 a + 2k_h b + c \right]$$

(16)

From Eqs. (10) and (15) it is to be noted that the class of estimators t_p' is preferable over the class of estimators t_p of two –phase sampling set up as t_p' is unbiased (up to first order of sample size) class of estimator of $\overline{Y}_h$ while the class of estimator t_p is biased.

7. Minimum variance of proposed class of estimators

It is obvious from the Eq. (16) that the variances of the proposed class of estimator t_p' depend on the value of the constant k_h. Therefore, we desire to minimize their variances and discussed them below. The optimality condition under which proposed class of estimators t_p' have minimum variance is obtained as

$$k_h = -\frac{b}{a}$$

(17)

Substituting the optimum value of the constant k_h in Eq. (19), we have the minimum variance of the class of estimators t'_p as

$$\text{Min.V}\left(t'_p\right) = \sum_{h=1}^{L} W^2_h \overline{Y}_h^{\,2} \left[f_1 C^2_{y_h} - \frac{b^2}{a} + C \right] \tag{18}$$

8. Efficiency comparison of the proposed strategy

It is important to investigate the performance of the proposed class of estimators with respect to the existing ones. We use the two natural population and one artificially generated population data set to justify the supremacy of the proposed strategy.

8.1 Empirical investigations through natural populations

The data set of two natural populations has been presented below.

- **Population I** (Source: Murthy [12], p. 228)

y: Factory **output** in thousand rupees, x: Number of workers in the factory, and z: Fixed capital of factory in thousand rupees.

The data consist of 80 observations which are divided into four strata according to the auxiliary variable z as: (i) $z \leq 500$, (ii) $500 < z \leq 1000$, (iii) $1000 < z \leq 2000$, and z > 2000 respectively for allocation of sample size to different strata, Proportional allocation is used.

Stratum 1 $(z \leq 500)$

$N_1 = 19, n'_1 = 11, n_1 = 5, \overline{Y}_1 = 2669.247, \overline{X}_1 = 65.15789$
$\overline{Z}_1 = 349.6842, C_{y_1} = 0.28363, C_{x_1} = 0.17153, C_{z_1} = 0.31299$
$\rho_{yx_1} = 0.81381, \rho_{yz_1} = 0.9364, \rho_{xz_1} = 0.9044$

Stratum 2 $(500 < z \leq 1000)$

$N_2 = 32, n'_2 = 17, n_2 = 8, \overline{Y}_2 = 4657.625, \overline{X}_2 = 139.9668$
$\overline{Z}_2 = 706.5938, C_{y_2} = 0.14366, C_{x_2} = 0.3169, C_{z_2} = 0.15457$
$\rho_{yx_2} = 0.8883, \rho_{yz_2} = 0.9259, \rho_{xz_2} = 0.8456$

Stratum 3 $(1000 < z \leq 2000)$

$N_3 = 14, n'_3 = 8, n_3 = 3, \overline{Y}_3 = 6537.214, \overline{X}_3 = 403.2143$
$\overline{Z}_3 = 1539.571, C_{y_3} = 0.06365, C_{x_3} = 0.20117, C_{z_3} = 0.18004$
$\rho_{yx_3} = 0.9295, \rho_{yz_3} = 0.9835, \rho_{xz_3} = 0.9366$

Stratum 4 $(z > 2000)$

$N_4 = 15, n'_4 = 9, n_4 = 4, \overline{Y}_4 = 7843.667, \overline{X}_4 = 763.2$
$\overline{Z}_4 = 2620.533, C_{y_4} = 0.08232, C_{x_4} = 0.22464, C_{z_4} = 0.14156$
$\rho_{yx_4} = 0.9787, \rho_{yz_4} = 0.9692, \rho_{xz_4} = 0.9454$

- **Population II** (Source: Koyuncu and Kadilar [13]).

y: Number of teachers, x: Number of students both primary and secondary schools, and z: Number of classes both primary and secondary schools. There are 923 districts in 6 regions (as: (i) Marmara, (ii) Agean, (iii) Mediterranean, (iv) Central Anatolia, (v) Black Sea, (vi): East and Southeast Anatolia) in Turkey in 2007 (source: The Turkish Republic Ministry of Education).

Marmara region

$$N_1 = 127, n_1' = 60, n_1 = 31, \overline{Y}_1 = 703.74, \overline{X}_1 = 20804.59$$
$$\overline{Z}_1 = 498.28, C_{y_1} = 1.25591, C_{x_1} = 1.46538, C_{z_1} = 1.115$$
$$\rho_{yx_1} = 0.936, \rho_{yz_1} = 0.97891, \rho_{xz_1} = 0.93958$$

Agean region

$$N_2 = 117, n_2' = 40, n_2 = 21, \overline{Y}_2 = 413, \overline{X}_2 = 9211.79$$
$$\overline{Z}_2 = 318.83, C_{y_2} = 1.56155, C_{x_2} = 1.64797, C_{z_2} = 1.14804$$
$$\rho_{yx_2} = 0.996, \rho_{yz_2} = 0.97624, \rho_{xz_2} = 0.96958$$

Mediterranean

$$N_3 = 103, n_3' = 50, n_3 = 29, \overline{Y}_3 = 573.17, \overline{X}_3 = 14309.3$$
$$\overline{Z}_3 = 431.36, C_{y_3} = 1.80307, C_{x_3} = 1.9253, C_{z_3} = 1.42097$$
$$\rho_{yx_3} = 0.994, \rho_{yz_3} = 0.98351, \rho_{xz_3} = 0.97655$$

Central Anatolia region

$$N_4 = 170, n_4' = 75, n_4 = 38, \overline{Y}_4 = 424.66, \overline{X}_4 = 9478.85$$
$$\overline{Z}_4 = 311.32, C_{y_4} = 1.90878, C_{x_4} = 1.92206, C_{z_4} = 1.47124$$
$$\rho_{yx_4} = 0.983, \rho_{yz_4} = 0.98296, \rho_{xz_4} = 0.96362$$

Black sea region

$$N_5 = 205, n_5' = 40, n_5 = 25, \overline{Y}_5 = 267.03, \overline{X}_5 = 5569.95$$
$$\overline{Z}_5 = 227.20, C_{y_5} = 1.51162, C_{x_5} = 1.52564, C_{z_5} = 1.14811$$
$$\rho_{yx_5} = 0.989, \rho_{yz_5} = 0.96434, \rho_{xz_5} = 0.96725.$$

The percentage relative efficiencies (PRE) the proposed class of estimators t_p' with respect to different estimators under their respective optimum conditions are shown below.

8.2 Empirical investigations through artificially generated population

An important aspect of simulation is that one builds a simulation model to replicate the actual system. Simulation allows comparison of analytical techniques and helps in concluding whether a newly developed technique is better than the existing ones. Motivated by Singh and Deo [14], Singh et al. [15] and Maji et al. [16] who have been adopted the artificial population generation techniques, we have generated five sets of independent random numbers of size N (N = 100) namely x_{1_k}', y_{1_k}', x_{2_k}', y_{2_k}' and z_k' (k = 1, 2, 3, ..., N) from a standard normal distribution with the help of R-software. By varying the correlation coefficients ρ_{yx} and ρ_{xz}, we have

generated the following transformed variables of the population U with the values of $\sigma_y^2 = 50,\ \mu_y = 10,\ \sigma_x^2 = 100,\ \mu_x = 50,\ \sigma_z^2 = 50$ and $\mu_z = 20$ as

$$y_{1_k} = \mu_y + \sigma_y \left[\rho_{xy} x'_{1_k} + \left(\sqrt{1 - \rho_{yx}^2} \right) y'_{1_k} \right]$$

$$x_{1_k} = \mu_x + \sigma_x x'_{1_k}$$

$$z_k = \mu_z + \sigma_z \left[\rho_{xz} x'_{1_k} + \left(\sqrt{1 - \rho_{xz}^2} \right) z'_k \right]$$

$$y_{2_k} = y_{1_k}$$

$$\text{and } x_{2_k} = x_{1_k}.$$

We have split total population of size N = 100 into 5 strata each of size 20 [i.e., $N_h = 20; (h = 1, 2, ..., 5)$] taking them sequentially and consider $n'_h = 12$ and $n_h = 8; (h = 1, 2, ..., 5)$ for the efficiency comparison of the proposed strategy.

The percentage relative efficiencies the proposed class of estimators t'_p with respect to different estimators (under their respective optimum conditions) are derived through the data set of the artificially generated population are obtained as:

9. Conclusion

From the construction of estimation strategy and efficiency comparison of the proposed methodology, following matters are noted.

1. Form **Table 1**, it is clear that the proposed class of estimators is at least 1% better than the existing one in estimating the population mean.

2. Similarly from **Table 2** it is found that the new estimator is at least 28% better than the existing one.

3. It may also be noted from **Tables 1** and **2** that the artificially generated population is homogeneous (the mean and variance of the respective variables are almost same for different strata) where the natural populations are heterogeneous (the mean and variance of the respective variables are different for different strata) in nature. Our suggested estimators performs with equal efficiency for both the types.

Estimator	PRE	
	Population I	**Population II**
$\left(\bar{y}_{st} \right)$	173.3608	192.951
$\bar{y}^{(h)}_{RP}$	101.1429	131.5654
$\bar{y}^h_{cs}$	118.3215	172.226

We use following expression to obtain the percent relative efficiency (PRE) of the proposed estimator t'_p with respect to different estimators as $\text{PRE} = \frac{V(\bar{y})}{\text{Min.V}(t'_p)} \times 100.$

Table 1.
PRE of the proposed estimator t'_p with respect to different estimators through data set of natural population.

Estimator	PRE
	Artificially generated population
$(\overline{y}_{st})$	179.623
$\overline{y}^{(h)}_{RP}$	128.256
$\overline{y}^{h}_{cs}$	154.879

We use following expression to obtain the percent relative efficiency (PRE) of the proposed estimator t'_p with respect to different estimators as $\text{PRE} = \frac{V(\overline{y})}{\text{Min.}V(t'_p)} \times 100$.

Table 2.
PRE of the proposed estimator t'_p with respect to different estimators through data set of artificially generated population.

4. The unbiased version of the proposed technique has been obtained which make the proposed class of estimators much more practicable.

Thus, it is found that the proposed estimation technique has addressed the problems of estimation through two-phase stratified sampling which may truthful for real life application where population is especially heterogeneous in nature and stratification is essential. Due to the benefits achieved by the new estimator, the survey statistician may be suggested to use it.

Author details

Partha Parichha[1], Kajla Basu[2] and Arnab Bandyopadhyay[1]*

1 Department of Mathematics, Asansol Engineering College, Asansol, India

2 Department of Mathematics, National Institute of Technology, Durgapur, India

*Address all correspondence to: arnabbandyopadhyay4@gmail.com

IntechOpen

References

[1] Sukhatme B. Some ratio type estimators in two-phase sampling. Journal of the American Statistics Associations. 1962;**57**:628-632

[2] Chand L. Some ratio type estimators based on two or more auxiliary variables [unpublished PhD thesis]. Ames, Iowa (USA): Iowa State University; 1975

[3] Kiregyera B. A chain ratio type estimators in finite population double sampling using two auxiliary variables. Metrika. 1980;**17**:217-223

[4] Kiregyera B. Regression type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrika. 1984;**31**: 215-226

[5] Tracy DS, Singh HP, Singh R. An alternative to the ratio-cum-product estimator in sample surveys. Journal of Statistical Planning and Inference. 1996; **53**:375-387

[6] Singh HP, Espejo MR. Double sampling ratio-product estimator of a finite population mean in sampling surveys. Journal of Applied Statistics. 2007;**34**(1):71-85

[7] Gupta S, Shabbir J. on the use of transformed auxiliary variables in estimating population mean by using two auxiliary variables. Journal of Statistical Planning and Inference. 2007; **137**:1606-1611

[8] Shukla D, Pathak S, Thakur NS. Estimation of population mean using two auxiliary sources in sample surveys. Statistics in Transition. 2012;**13**(1):21-36

[9] Choudhury S, Singh BK. A class of chain ratio–product type estimators with two auxiliary variables under double sampling scheme. Journal of the Korean Statistical Society. 2012;**41**: 247-256

[10] Parichha P, Basu K, Bandyopadhyay A, Mukhopadhyay P. Development of efficient estimation technique for population mean in two phase sampling using fuzzy tools. Journal of Applied Mathematics, Statistics and Informatics. 2017;**13**(2):5-28. DOI: 10.1515/jamsi-2017-0006

[11] Bandyopadhyay A, Singh GN. Predictive estimation of population mean in two-phase sampling. Communications in Statistics: Theory and Methods. 2016;**45**(14):4249-4267. DOI: 10.1080/03610926.2014.919396

[12] Murthy MN. Sampling Theory and Methods. Calcutta: Statistical Publishing Society; 1967

[13] Koyuncu N, Kadilar C. Family of estimators of population mean using two auxiliary variables in stratified sampling. Communications in Statistics: Theory and Methods. 2009;**38**: 2398-2417

[14] Singh S, Deo B. Imputation by power transformation. Statistical Papers. 2003;**4**:555-579

[15] Singh S, Joarder AH, Tracy DS. Median estimation using double sampling. Australian & New Zealand Journal of Statistics. 2001;**43**(1):33-46

[16] Maji R, Singh GN, Bandyopadhyay A. Estimation of population mean in presence of random non-response in two-stage cluster sampling. Communications in Statistics: Theory and Methods, ISSN: 0361-0926. 2018. DOI: 10.1080/03610926.2018.1478101

Chapter 4

Methods of Russian Patent Analysis

Dmitriy Korobkin, Sergey Vasiliev, Sergey Fomenkov and S.G. Kolesnikov

Abstract

The article presents a method for extracting predicate-argument constructions characterizing the composition of the structural elements of the inventions and the relationships between them. The extracted structures are converted into a domain ontology and used in prior art patent search and information support of automated invention. The analysis of existing natural language processing (NLP) tools in relation to the processing of Russian-language patents has been carried out. A new method for extracting structured data from patents has been proposed taking into account the specificity of the text of patents and is based on the shallow parsing and segmentation of sentences. The value of the F1 metric for a rigorous estimate of data extraction is 63% and for a lax estimate is 79%. The results obtained suggest that the proposed method is promising.

Keywords: patents, information extraction, SAO, ontology, prior art patent search

1. Introduction

From year to year, the number of patents and patent applications is increasing. The escalating applications flow, and more than 20 million world set of granted patents (from 1980 to 2015) increase the time that patent examiners have to spend to examine all incoming applications. Also, automation development of inventions has been gaining momentum, and computer-aided invention (CAI) systems are used to search for new technical solutions. The success of such systems largely depends on the completeness of the ontologies of the subject areas and the fullness of the various knowledge bases that allow generating new technical solutions.

The task of prior art patent search and information support for new technical solution synthesis can be seen as the task of extracting the subject-action-object (SAO) semantic structures. Patent claims are considered to be a direct source of data for retrieval. They express the essence of the invention, and therefore it is of the greatest interest for the SAO extraction.

In the paper [1], authors proposed a methodology to solve the problem of prior art patent search, consisting of a statistical and semantic analysis of patent documents, machine translation of patent application, and calculation of semantic similarity between application and patents on the base of subject-action-object (SAO) triples. On the step of the semantic analysis, the authors applied a new method for building a semantic representation of SAO on the base of meaning-text theory [2]. On the step of semantic similarity calculation, the authors compare the SAOs from

the application and patent claims. The results of checking the Russian- and English-language version of the semantic analyzer showed a low recall value for Russian patents—this is due to the complexity of Russian grammar.

Current approaches to extracting structured data and natural language processing (NLP) tools are poorly oriented to work on an array of Russian patents. In this regard, it is necessary to develop new efficient methods for extracting data from Russian patents.

The task of prior art patent search and information support for new technical solution synthesis can be seen as the task of extracting the subject-action-object (SAO) semantic structures. In particular, the elements of the design of a technical object and the relationship between them are considered as the source of information support.

Claims are considered to be a direct source of data for retrieval. They express the essence of the invention and are based on a description; therefore, it is of the greatest interest for the extraction of technical information. In this case, the object of analysis is an array of Russian-language patents.

This paper analyzes the application of the current NLP tools and data extraction approaches in the context of the current task. The article offers a new method of extracting SAO structures and constructing a graph of structural elements as part of solving the problem of prior art patent search and information support for the synthesis of new technical solutions.

2. Research background

Various systems [3–6] are known for processing English-language patents, including the use of the SAO formalism to extract various concepts. In [6], a tree syntax analysis was applied with a separate identification of the subject, the action, and the object to improve the quality of parsing. The authors state the following averaged values of precision and recall: 0.8058 and 0.8446, respectively. At the same time, the project was implemented in the GATE system, which is poorly suitable for the Russian language. In [7], the SAO structures are extracted based on the rules using the Stanford parser software, but there are no ready-made models for the Russian language. Papers of [4, 8] process patents using linguistic markers (specific verbs and nouns) and lexical-syntactic patterns, while [8] notes the need to study the structure of patent documents to improve the quality of data extraction. The emphasis is mainly on rule-based systems, since for statistical analysis systems, a lot of markup data is obviously needed. This causes additional development difficulties.

The logic of constructing a claim of the invention of the European type, applied in the Russian Federation, includes statement of the formula, starting with the generic term, and reflecting the purpose of the invention; after the expression, "including," "containing," or "consisting of," sets out the restrictive part of the patent claim, including the features of the invention, coinciding with the signs of the closest analogue and then after the phrase "characterized in that," its distinctive part is introduced, including the features that distinguish the invention from the closest analogue [9]. The formula is written in one sentence.

The specific style of writing a patent claims complicates the work of existing parsers, highlighting the following problems:

- The length of the sentences of the claims is excessively long.

- The claims are described predominantly with noun phrases, often with a chain of definitions and participial constructions.

- The availability of specific words and constructions can affect the correct operation of NLP tools.

Considering the system for extracting structured data from Russian-language texts, the development of Yandex Tomita-parser [10] is most often distinguished. Extraction takes place according to the rules, written in an extended context-free grammar.

In [11], the authors emphasize the advantages of Tomita-parser in the tasks of extracting named entities; the declared value of the F1 metric is 78.13%. In [12], this tool is successfully used to extract metadata from full-text electronic Russian-language publications with correct extraction of 86.7% of metadata. In [13], the Tomita-parser is used to extract from the patents SAOs describing the so-called technical functions.

However, due to the specifics of writing the claims on the invention, there are difficulties in using the Tomita-parser:

- Homogeneous subjects are not extracted using coordination mark of subject and predicate [10], since coordination of the whole bundle is not taken into account.

- Writing grammars for all possible parse trees with regard to the free word order in the Russian language is a laborious procedure.

- Incorrect identification of actants (erroneous determination of the cases of words, the need to take into account the valencies of the verb).

Despite the expressive power of context-free grammars and the Tomita-parser tool, the question of organizing the effective extraction of SAO structures from the claims remains open.

Of the available systems of syntactic and semantic-syntactic analysis, for which there is an opportunity to work with the Russian language, we can distinguish Link Grammar Parser [14], MaltParser [15], and UFAL UDPipe [16].

Link Grammar Parser is an open-source grammar parser. In [17], a system is described for extracting technical functions in SAO format from English-language patents using context-dependent grammar on Link Grammar Parser markup. At the same time, the results of experiments are presented (with an extraction precision of up to 0.85 and recall up to 0.78). The quality of work of Link Grammar Parser with the Russian language is the open question.

MaltParser is a tool for working with dependency trees, allowing to train a model on the annotated corpus and build trees for new data based on it. It is possible to train an own model on the annotated corpus of the Russian-language SynTagRus, openly laid out at the Universal Dependencies project [18]. The prepared morphological markup in the format CoNLL-U [19] is supplied to the analyzer input, which is supplemented in the process of analysis by the arrangement of vertices and types of links. In [20], estimates of the quality of semantic analysis of tools based on MaltParser are given; the value of the F1 measure reaches 71.0%.

UDPipe is a trained tool for tokenization, tagging, lemmatization, and dependency parsing. The project is open and actively developing. Trained models are available for a variety of languages, including Russian. The format of the input and output data is presented in the CoNLL-U markup. For the model trained on the SynTagRus package, the markup accuracy (on raw text) for all tags, as well as for the labeled attachment score (LAS) and unlabeled attachment score (UAS) metrics, is 93.2, 85.3, and 87.9%, respectively.

One of the main obstacles to the use of these systems is the excessive length of the sentences of the claims and, in particular, the considerable distance of homogeneous members from the binding predicate. The latter has a significant impact on the completeness of the extraction of objects.

Omitting the issues of training models, data selection, analysis tools, morphologists, and other things, in general, it can be concluded that the syntax and semantic-syntactic markup tools (as reviewed) work extremely unstable on long texts. It is impossible to use direct analysis results for efficient data extraction. At the same time, it is this structure of the description that prevails in the texts of the patent claims.

A possible solution in reducing the complexity of parsing text is its segmentation. The work [21] presents an algorithm for the segmentation of the text of the claims, based on the specific markers of the division of sentences. This approach reduces the length of the analyzed segment of the proposal but is not sufficient for a full analysis due to the possible absence of signs of separation in complicated sentences.

One of the most extensive works on the full segmentation of sentences in Russian texts can be considered the works of T. Yu. Kobzareva. The book [22] describes the elements of grammar and algorithmic solutions of the system of shallow parsing. In general, the system is set up to analyze the proposals of the Russian language, but it is rather cumbersome.

In [23], on the basis of the analysis of the existing parsers, the author builds the analyzer version, the segmentation module of which, by its functionality, closely matches the task at hand.

There are various approaches to segmentation, for example, taking into account the valences of verbs, using patterns of bases, etc. However, these solutions are universal with respect to texts and therefore do not have specific rules for the analysis of the claims of the invention and in some cases are somewhat redundant.

When building own segmentation and data extraction modules (low-level approach), it is necessary to have reliable data on the morphology of the sentence: analysis of the form of incoming words and their grammatical categories (gender, case, etc.). Among the available Russian-language analyzers that can be identified are TreeTagger, MyStem, TnT, pymorphy2, and FreeLing. A number of works [24–26] are devoted to the comparison of morphological analyzers of the Russian language. In the works [25, 26] on the quality of determining the parts of speech (POS), TreeTagger is best distinguished, but they do not take into account the work of the Yandex MyStem tool. In [24], already taking into account MyStem, the latter was chosen as the best tool with the F1 metric value of 94%. At the same time, an experiment is being conducted to improve the quality of parsing by combining tools (correcting MyStem with pymorphy2 with an increase in F1 to 95%). However, the full use of the analyzer MyStem is difficult due to the issuance of grammatical signs in the form of an unordered list of alternatives. At the same time, there are no attributes for the correct choice of a single grammatical form; the alternatives simply follow in alphabetical order. The TreeTagger morphoanalyzer is devoid of such a feature—the grammar is issued without alternatives. The disadvantage of TreeTagger is that its speed is somewhat lower than the speed of processing MyStem on large texts.

Having considered certain aspects of the processing of Russian-language texts, we summarize the problems that arise when extracting structured data (SAO structures) from texts of Russian-language patents when solving the problem of prior art patent search and information support for automated invention:

- The claims are written in a specific sublanguage of patents that impede text analysis (specific terminology).
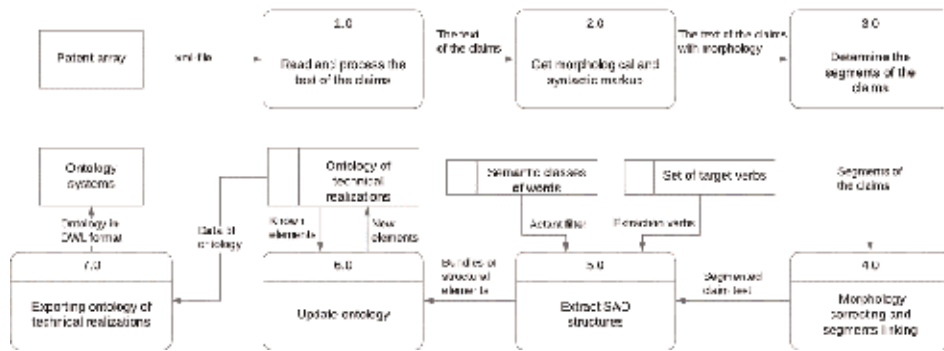
Figure 1.
DFD chart of the proposed system.

- The structure of the claims is poorly suited for parsing by ready-made syntactic analyzers (given the prevalence of sentences and the presence of multiple participle clauses).

- The structure of the claims is not fully decomposed by simple heuristics.

- The available NLP tools will certainly be mistaken in the analysis, so relying on one tool in the processing of complex structures is not always rational.

The presented limitations make it difficult to create an efficient data extraction system from patent claims using publicly available NLP tools (without considering complex commercial or closed systems).

However, given that the texts of patents are written in a template (not only in the structure of sections but also in the structure of phrases), it is possible to combine NLP tools for mutual correction of analysis errors; the authors suggest the following approach in processing the text of the claims, including the steps:

- Segmentation of the patent claims based on shallow parsing

- Organization of the extraction of SAO structures based on the valence bond theory

- Post-processing of SAO structures with linking concepts to a common repository

Next, we will consider building a system based on the findings.

The main stages of processing and data flows of the proposed system are presented in **Figure 1**.

In this research, the stages of the ontology organization are not considered (actions 6.0, 7.0). The end result will be the extraction of SAO structures and the construction of a graph of the elements of the invention.

3. Method description

The unit of extraction is the predicate-argument construction in the form of the SAO structure, which semantically describes the structural elements of the device and the connections between them.

The following instruments of NLP in the experimental mode are used:

- UDPipe (tokenization, tagging, lemmatization, and syntactic analysis)

- MyStem (lemmatization, correction of morphology)

- pymorphy2 [27] (declination of word forms)

- Chanker of noun phrases of Russian [28] (to define the boundaries of actants)

The ultimate goal of extracting data is to build a related graph of structural elements of the invention, reflecting the structure of the invention itself. Let us consider in more detail each of the stages.

3.1 Segmentation of patent claims

On the basis of a well-established patent language, the authors made the assumption that it is possible to identify certain patterns of segments in a patent claim and that shallow parsing with a number of heuristics will be enough to successfully extract the necessary structures.

The segment of the independent claim will be called the element of the sentence, which is indicated by punctuation (comma and/or semicolon). In the Russian language, such segments can act as simple main sentences and be subordinate clauses, participle and verbal adverb phrases, isolated definitions, etc. In addition, one should take into account the logical division of the claim into the restrictive and distinctive parts (if such separation exists), usually demarcated by the construction of "characterized in that."

The purpose of segmentation is to identify the linear structure of the proposal of the claims for the subsequent extraction of SAOs based on the analysis of the morphological and linear combinatorial characteristics of the text. Let us explain it with the example of the text of the patent "Pressure sensor," the syntactic analysis of which is presented in **Figure 2**.

In parentheses are marked linear numbers of segments without taking into account the demarcation design "characterized in that." At the same time, the verb forms without specifying the role in the sentence are highlighted with a double line.

The sentence includes generic term (1); participle clause (2, 4, 8); subordinate-attributive (3, 5); and simple two-part (7, 9). The tasks of segmentation include the identification of segments by composition and the definition of the relationship between the main and dependent words. In the future, this information is necessary for the direct extraction of SAOs.
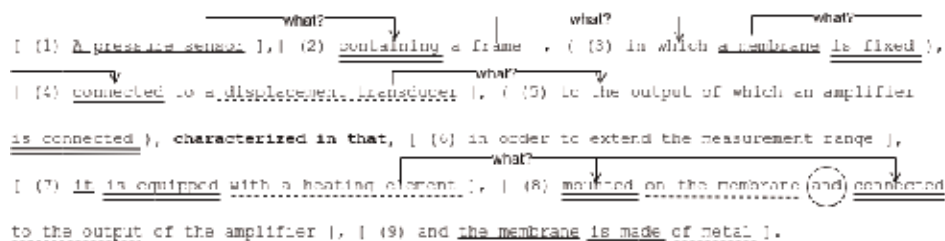

Figure 2.
Parsing sentence.

Type of segment	Symbol	Example of a segment
Noun phrase	N	"contact electric current modulator"
Participle clause	V_N	"containing vibration transducer"
Independent segment	N_V_N	"it is equipped with an additional vibration converter"
Isolated attribute	A0	"identical to the main"
Rather independent segment	(N_) which_V_N	"which input is the input of the device"
Dash instead of a predicate	N_-_N	"and the cathode through resistance—with the anode"
Comparative turns	AV	"less"
Subordinate clauses	SC	"that they have sufficient fatigue strength"
Verbal adverb phrase	V<C>	"preventing the deformation of the first destructive elements"
Service segments	SEPARATOR	besides; and; and also; at the same time
	NAMED	"called decaying elements"
	PURPOSE	"for the purpose of range extension of measurements"

Table 1.
Types of segments of the test sample claims.

To identify the nature of writing segments and formalize the definition of their types, a set of 14 main claims was taken, mainly from different sections of the international patent classification (IPC). The total number of segments in the sample was 187 elements.

After a detailed manual analysis, the following conditional segment types were derived (see **Table 1**).

At the same time, the symbol for the segment was formed from acronyms, taking into account the basic structure of the segment, i.e., the POS and/or elements that characterize it: noun (N); verb (V); adjective (A); "which" is a demonstrative pronoun in subordinate clauses; adverb (AV); and others. The legends of the segments in the future act as markers in the system.

Apparently from the results of the analysis, all the segments met a finite number of fairly typical sentence constructions. Of course, it is not possible to thoroughly analyze the entire array of patents manually. Therefore, the authors made the assumption that the number of types of segments (in terms of the overall structure of writing) of a larger mass of the claims is also limited.

Since the segmentation is carried out in the context of the task of extracting predicate-argument structures, the elements of this extraction are partially embedded in the process in the form of highlighting nonsignificant (service) segments and the target semantic classes of verbs.

Semantic classes of verbs are needed to find to SAO target ligaments, potentially describing constructive elements and the connections between them. The samples of such verbs are presented in **Table 2**.

The overall sequence of operations for the segmentation of the claims is represented by the activity diagram in **Figure 3**.

1. Text preprocessing includes the following steps:

   • Deletion of introductory words: "for example," "at least," "however," etc.

   • Removal of brackets with all contents: for example, references to figures "(1), (12)"

Semantic class	Verbs
Existence of a structural element	contain (on); equip; include (between; in; with; in;); have; supply; enter (into); consist (from)
Inter-element couplings	set (on; as); connect (to); fasten (in; on); place (in; on)

Table 2.
Semantic classes of verbs.
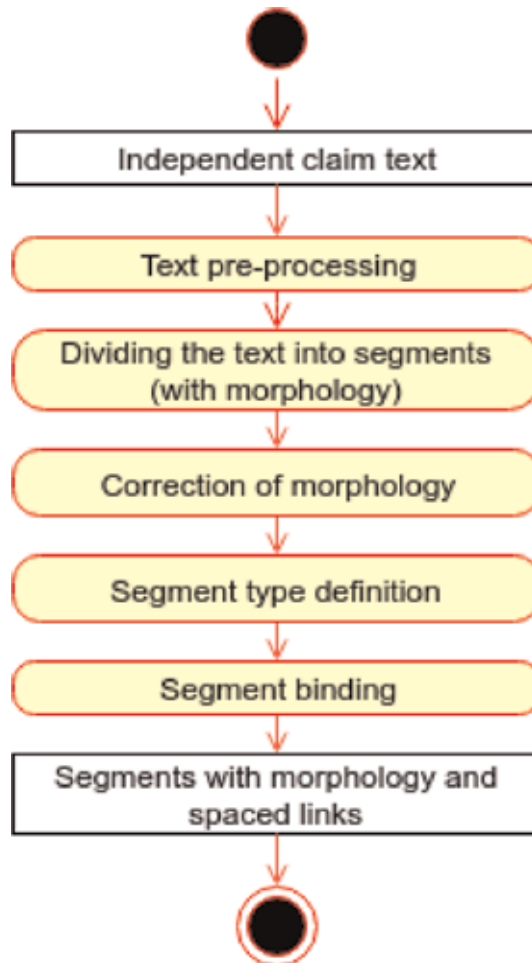
Figure 3.
The segmentation process of the claims.

- Deletion of the claim number at the beginning of the text ("1. Magnetic gear … ")

- Removal of html tags (<. *?>) in case of reading from xml documents

- Addition of spaces to slashes (to correct the morphology to "and/or")

- Removal of multiple spaces and duplicate commas

2. The process of obtaining initial segments is associated with sequential fragmentation of the text by punctuation marks with auxiliary treatments. A

specific feature of the stage is the final fragmentation based on the results of the syntactic analysis obtained after using the UDPipe tool. Tokenization is also performed by this tool. This decision is justified by the need to take into account the context for better analysis: submitting partially fragmented text to the analyzer input (based on strong signs of separation—e.g., different parts of the claims or a semicolon), but without carrying out fragmentation on commas, it is possible to receive more adequate syntactic marking; analysis of segments completely isolated from each other potentially also leads to errors.

The division of the text into segments includes the steps:

- Division of the text according to the pattern, "characterized in that," while the segment is marked with a conditional part number: (0, all segments belong to one part (claim without division); 1, segment of the restrictive part; 2, segment of the distinctive part)

- Division of the text by a semicolon

- Insertion of surrogate references to formulas and complex tokens (e.g., "S = W = 0.00042λ") to reduce segmentation errors and morphological analysis

- Obtaining the primary morphology from the UDPipe tool and segmentation by commas

3. The next step is the process of correcting morphology. If a segment token refers to parts of a verb, adjective, or adverb speech, then a part of speech is checked using MyStem. Moreover, if the result does not coincide with the initial one, then the morphological information of the token is updated taking into account the conversion of tags from MyStem into UDPipe format (CoNLL-U).

One of the fundamental stages of segmentation is the determination of the type of segment (see **Table 1**). The type of segment allows you to introduce certainty in the use of certain heuristic processing in the future. This problem is solved in two steps:

a. Determining a segment pattern

b. Determining the type of segment by pattern using a finite state machine (FSM)

The definition of a segment pattern is based on finding out its main composition, which is based on the priority of the presence of POS. The dictionary of symbols in the template is presented in **Table 3**.

A generalized algorithm for determining the pattern of a segment by the POS found in it is shown in **Figure 4**.

An example of the output templates and types for the segments is an example from **Figure 2**.

Segment (8): "mounted on a membrane and connected to the output of an amplifier" ("установленным на мембране и подключенным к выходу усилителя" in Russian).

Parts of speech (ru): [VERB ADP NOUN CCONJ VERB ADP NOUN NOUN]
Template: "V <P> _N_CC_V <P> _N"
Type: "V_N"

Symbol	Part of speech	Note
N	Noun	Index 0 means accusative or a nominative case
V	Verb	Possible litters: <P> participle; <F> finite verb; <C> adverbial participle; <I> infinitive; <U> undefined
A	Adjective	Index 0 means the top of the sentence (indirect attribute)
SC	Subordinate conjunction	—
CC	Coordinative conjunction	—
PR	Pronoun	Index 0 means the nominative case Direct replacement of the word "which" is possible
AV	Adverb or numeral	—
X	Formula	For the case of explanation of the formulas
—	Dash	Verb skip or explanation

Table 3.
Template dictionary.

The definition of a segment type by the template is implemented using an FSM. The principle is to search for elements of the segment template and search for specific sequences of POS corresponding to the types of segments.

4. Let us describe the process of binding of segments. The main binding mechanism in the Russian language is coordination according to gender, number, and case. Therefore, the segmentation is carried out on the data of the extracted morphology, and the task is reduced in finding this agreement between the words in the sentence. At the same time, segment types limit the set of binding rules.

   Proceeding from a linear logic of the description of the formula, it is possible to a connection of segments in one straight pass (from left to right). An exception is the case of a rupture of the sentence by an enclosed turnover. As a rule, the second part of a broken sentence begins with a finite verb. When finding such a segment, it is necessary to find the closest missing segment of the noun group to the left. The assembly of such segments precedes the main straight passage.

   Next, each segment type is processed. It is necessary to explain the mechanism of binding to the predicate forms of homogeneous members and constructions:

   ```
   <(1) AC voltage regulator, containing (2) a discrete-action
   regulator, connected between the input and output terminals,
   (3) an interval converter into a code, whose information input
   is connected to the output terminal, a synchronizing input – to
   the input terminal, ..., as well as (4) a logic device>.
   ```

   The generic term is "AC voltage regulator" (1). The participle "containing" combines the enumerated structural elements of the device ("regulator" (2), "converter" (3), and "logic device" (4)). However, some homogeneous members (in the example it is "logic device" (4)) may be located far away from the describing segment (participle "containing").

To solve this problem, firstly, a dictionary of predicates is introduced (see **Table 2**), which is semantically pointing to the composition of the object (e.g.,
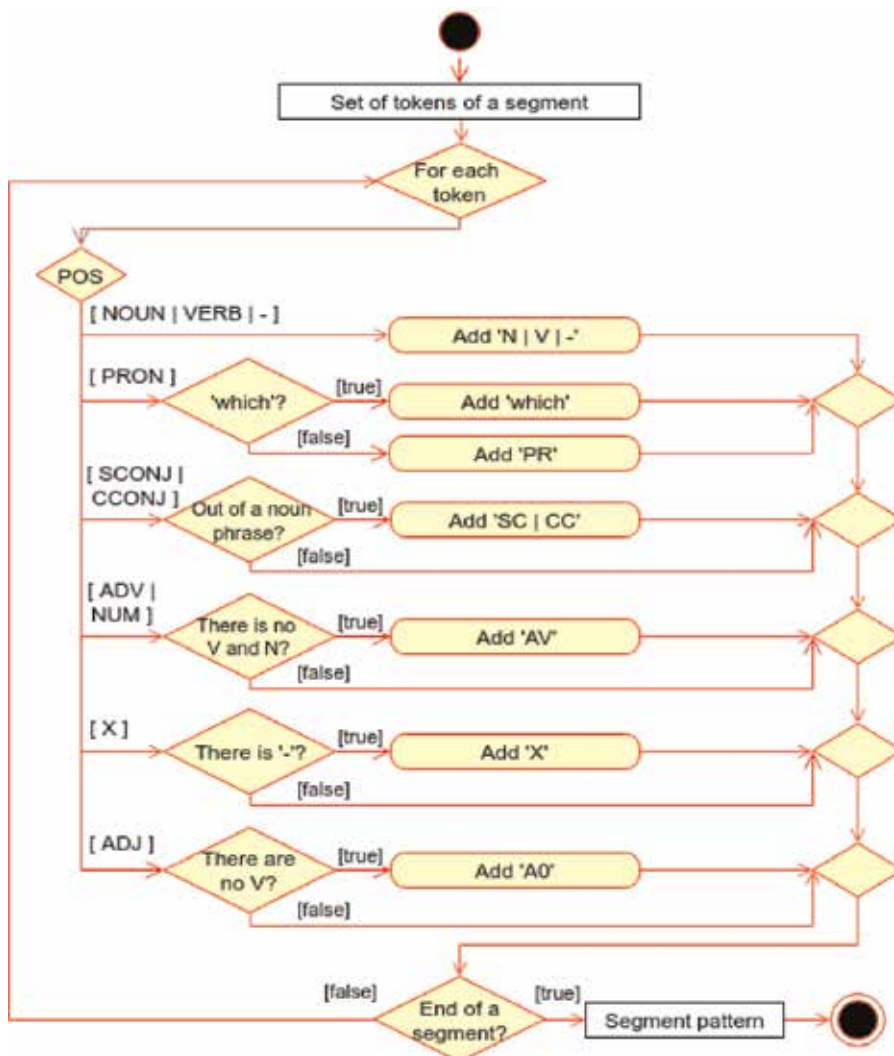
Figure 4.
Segment pattern generation sequence.

"contain," "include," "have," etc.); secondly, the mechanism of binding to predicates is introduced.

Anchor points (predicates) are tracked as structures:

1. Last compositional verb (LCV), searched by dictionary: LCV = (<segment_id>, <verb_id>).

2. Last active verb (LAV), remembers the last predicate in the corresponding types of segments, since homogeneous terms can follow: LAV = (<segment_id>, <verb_id>, <noun_case>). The LAV structure stores information on the case of a noun following the selected verb for anchoring correction.

Example of binding to predicates:

```
 containing[LCV] an amplifier (Nom/Acc case), connected[LAV] to
the source (Dat case) of the power supply, filter (Nom case)...
```

A brief description of each type of segment is presented in **Table 4**.

The result of segment linking for an example (**Figure 2**) is shown in **Figure 5**.

A pseudo-vertex (ROOT) is introduced on the graph, to which all independent, initial, or unrecognized segments are attached. According to the graph, it can be noted that the participial revolutions (2, 4, 8) are linked to the parent segments. Segments with the demonstrative pronoun "which" (3, 5) are also attached; the type of the connection "spec" is marked. In general, the entire convolution is performed correctly.

Segment type	Heuristic rules
N	The first segment of the entire claims is declared independent Otherwise, a sequential attempt to bind to LAV and LCV
V_N	In the presence of participle phrase, which starts with a co-coordinating conjunction, an attempt is made to bind to LAV and LCV. Check on the construction of V (plural) N (singular): in the presence of LCV, this segment will be homogeneous. Binding of the first predicate to the noun in the agreement of gender, number, and case. Rewriting LAV and LCV structures
N_V_N	The default is an independent segment. Check for a possible type determination error by the presence of V<P> at the beginning of the segment followed by linking to the main word by agreement
A0	In the case of the beginning of a segment with an adjective, a binding is made to the noun in the previous segments as agreed A segment can be declared independent if it starts with a noun and ends with the adjective A0
N_which_V_N, which_V_N	The specifying segment contacts a noun in the previous segment under the approval of a gender, number, and case of the word "which"
N_-_N	A segment with a missing predicate is bound to the previous segment if it is not a description of the formula
AV/SC	Adverb or comparison/complex sentence. The default binding to the previous segment
V<C>	The verbal adverb phrase is looking for a point of attachment in the form of a segment with a finite verb
PURPOSE	Segments describing the purpose of constructions depending on the form of the record are tied to the previous or next segment ("intended …" or "for the purpose of …")
NAMED	The names of the various elements have a mechanism for linking the participial phrases ("called …")

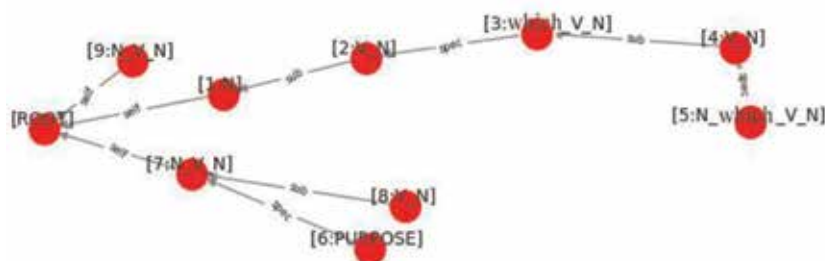**Table 4.**
Heuristic rules for segments.

Figure 5.
*The graph of segments for an example (**Figure 2**).*

3.2 Extracting SAO primary structures

The extraction of predicate-argument constructions is based on the results of segmentation and valency of the target verbs (see **Table 2**). The valence of the verb determines the number of possible arguments. It is enough to define the "subject" and "object" of the structure extracted.

The subject can be determined by the following cases (see Example 1), by the nominative case (1) and by the word of the participle being defined (2), or it can be completely absent (3), implying a generic term.

Example 1. The definition of the subject in the segments:

1. The magnetic gearbox is equipped with a rotor …

2. … magnetic gearbox, equipped with a rotor …

3. … characterized in that … , a rotor is introduced, …

The definition of the object is somewhat complicated due to the lack of explicit universal expression in the predicate arguments. The role of an object in a predicate corresponds to a specific case, which can be formed using prepositions.

Consider the cases of determining the object from Example 2: accusative case without preposition (1), dative case formed with the preposition "to" (2), and several actants with prepositions (3); in the latter case, the preposition "through" generates an adverb (connected as?—"through a non-magnetic ring"), and the target actant is a combination with the preposition "with" (connected with what?—"with the body").

Example 2. The definition of the object in the structure:

1. including shaper

2. connected respectively *to* the output of the amplifier

3. connected *with* the body *through* a non-magnetic ring

It should also be taken into account that the predicate can be expressed by a verb with a compound preposition (e.g., "made in the form of"). In addition, information on binding to the predicate can already be partially determined at the end of the segmentation.

To account for the valences of the target verbs, the following structure is introduced:

```
verb (obj_valence, add_valence, entry)
where verb - the predicate lemma in the dictionary;
entry - the position of actants relative to the predicate (0 - only
the sequence subject - action - object; 1 - any sequence of subject
/ object; 2 - only the sequence object - action - subject);
obj_valence - possible object valencies, includes the structure:
  obj_valence (
  before (
    type,
    mandatory,
    tokens
    ),
```

```
case
)
    where before – the sequence of words (prepositions) to actant,
may be absent,
    type – the type of the sequence (preposition or set of words),
    mandatory – obligatory presence of a preposition before the
construction
    tokens – sets of tokens in sequence,
    case – case with valence;
add_valence – additional object valencies, similar in structure
to the obj_valence structure, with the exception of adding the
type of additional valence.
```

To clarify the structure, consider Example 3: description of the valence of the verb "contain."

The valency of the object (obj) is represented by the accusative case Acc, and the nominative case Nom is added because of the possibility of a false definition of the case by the analyzer; however, there are no prepositions before the actant (structure before = none). Additional valencies (add) contain the valence with the locative case Loc and the obligatory preposition "on" or "at" in front of the actant.

The additional valence generates the auxiliary type ON (see all options in **Table 5**).

Accounting for additional valencies is necessary to eliminate the false definition of the subject and object. In this case, the object is defined by the nominative case in full-length segments and the anchor point in the subordinate segments.

The process of predicate-argument structures extraction in a generalized form is presented in the activity diagram (see **Figure 6**).

When detected in the segment of the target verbs (see **Table 2**), the sequence of currents to the right and left of the verb extends to the left and right parts, respectively. For full-length segments (types "N_V_N," "N_which_V_N," "which_V_N," and "N _-_ N" from **Table 1**), the right and left parts are formed in the same segments, and for subordinates (type "V_N"), the left part is replaced and forms the found subject connecting.

Next comes the key stage of extraction: the search for the structures of the object and subject in the selected parts of the segment according to the predicate valencies. The order of searching for a subject and an object is determined by the possible position of actants indicated in valency (the entry field). The sequence of determining the actants in the parts of the segment for any position of the subject and object is presented in **Figure 7**.

The search for valencies in the definition of a subject and an object in parts of a segment is carried out using the token-marking mechanism. The following label sets are used:

- For the subject (P-SBJ, N-SBJ, I-SBJ, and A-SBJ)

- For an object (P-OBJ, N-OBJ, and I-OBJ)

- For auxiliary valencies (P-ON, P-VIA, N-ON, N-VIA, N-BY, N-WHOM)

The label prefix corresponds to the functional purpose of the token: preposition (P); noun (N), and adjective (A), the top of the noun group; and inside (I), the entry of a token into a noun group.

Type of valence	Preposition before actant	Case of actant
ON	On/at	Prepositional case
VIA	Through	Accusative case
BY	—	Instrumental case
WHOM	—	Accusative case

Table 5.
Generated types of additional valencies of verbs.
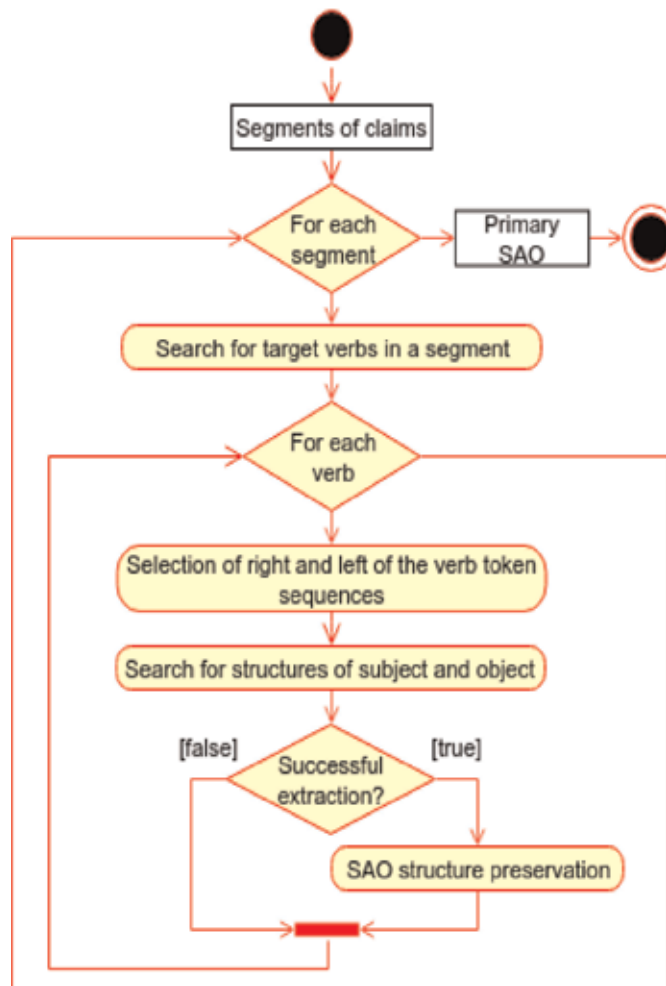
Figure 6.
The general algorithm for extracting SAO structures.

Marks in parts of the segment are placed for the object according to the predicate valences and for the subject, for the nominal case of a noun (or less often an adjective). If there is a mandatory preposition in valency, the occurrences of the specified preposition are first sought; in case of failure, the search is interrupted. The search for additional valences for an object is more priority than the search for the valence of the object itself, i.e., starts first.

Let us explain the presented methodology of SAO extraction on the example of a phrase: "[gasket] contains [at the other end one element]" ("[прокладка] содержит
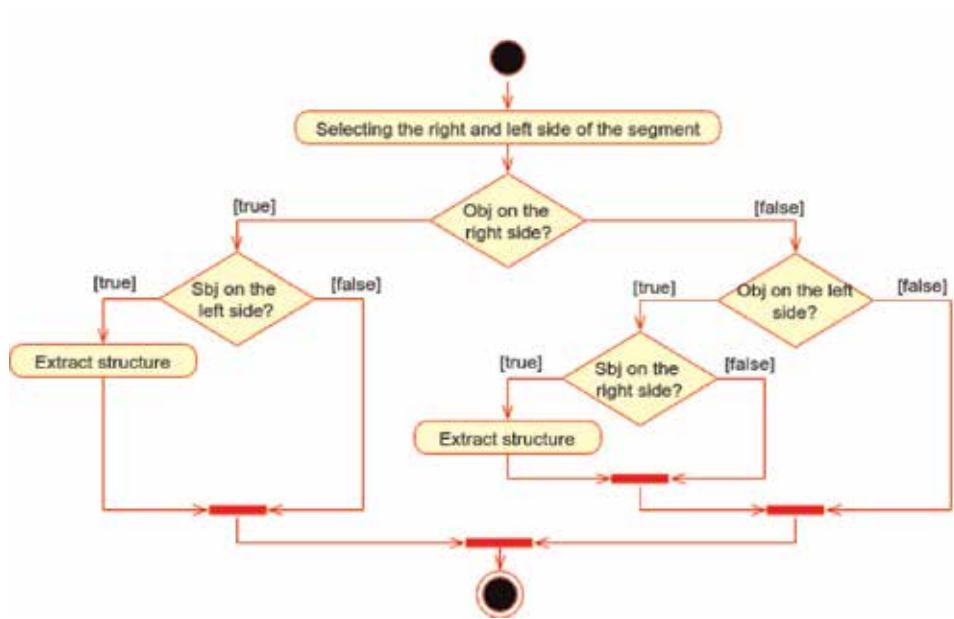
Figure 7.
The sequence of defining the actants.

Token	Lemma	POS
"gasket" ("прокладка" in Russian)	прокладка	NOUN
"contains" ("содержит" in Russian)	содержать	VERB
"at" ("на" in Russian)	на	ADP
"other" ("другом" in Russian)	другой	ADJ
"end" ("конце" in Russian)	конец	NOUN
"on" ("один" in Russian)	один	NUM
"element" ("элемент" in Russian)	элемент	NOUN

Table 6.
Morphology for each token.

на [другом конце один элемент]" in Russian). In this case, brackets indicate the boundaries of noun groups. A brief morphological analysis for each token is presented in **Table 6**.

Step 1. Search for target verbs

The target verb is the "contain" token, the valencies of which are described in Example 3.

Step 2: Select the right and left sequence of tokens from the verb

Left side (left part, l_part): "gasket"

Right part (right part, r_part): "one element at the other end"

Step 3: Search for structures of subject and object

A single sequence of subject and object is possible relative to the predicate "contain": subject-action-object (the entry field). Therefore, the search object is carried out only in the right part (r_part).

Step 3.1: Search for additional valencies

The predicate has additional valencies (general search pattern: ["on/at" Noun < Loc>]), for the object they are searched for first.

There is a search for a mandatory preposition "on/at," it is present on the right side of the segment and is marked (1). After that, the labeling of the additional valence continues: it is necessary to find a noun with the prepositional (Loc); it is also present (the token "end") and marked (2). Since the valence is actually found, it is necessary to label homogeneous members (from the preposition to the noun within the noun) with a marker with the prefix "I" (3).

Final marking after searching for additional valencies:

Token	At	The other	End	One	Element
Marker	P-ON (1)	I-ON (3)	N-ON (2)	—	—

Step 3.2 Search for object valence
The valency of the object contains only one position:

```
'obj':[{
    'before': None,
    'case':[ 'Acc','Nom'],
}]
```

In this case, before the actant, there are no prepositions. It is enough to find a noun, pronoun, or adjective in the necessary case (accusative (Acc) or nominative (Nom) case); and the accusative case will be sought first since it is listed first. Such an anchor point is the token "element," and it is marked with the label of the vertex of the noun group N-OBJ (1). As valency is found, it is necessary to complete the marking of a noun phrase of the vertex. In the sentence, there is the only token "one" (the marker of a noun phrase of a token "one" (2) matches a marker of a noun phrase of vertex "element"). Other markers (additional valencies) are not rewritten.

Final marking after the search of a valency of an object:

Token	at	the other	end	one	element
Marker	P-ON	I-ON	N-ON	I-OBJ (2)	N-OBJ (1)

Thus, an object is found, and it is possible to pass to the search for the subject.
Step 3.3 Search for subject valency
Subject search is carried out in the remaining left part. Since this is a full segment, as a subject it is necessary to find a noun, pronoun, or numeral in the nominative case. In the left part, there is only one token, "gasket," and it satisfies the search criteria, marked as N-SBJ.
Final marking of the right side of the segment:
Step 3.3 Search for subject valency
Subject search is carried out in the remaining left part. Since this is a full segment, as a subject it is necessary to find a noun, pronoun or numeral in the nominative case. In the left part there is only one token - "gasket", and it satisfies the search criteria; marked as N-SBJ.
Final marking of the right side of the segment:

Token	gasket
Marker	N-SBJ

Subject	"gasket" [N-SBJ]
Action	contain
Object	"at the other end **one element**" → "one element" [P-ON I-ON N-ON **I-OBJ N-OBJ**]

Step 4. Formation of predicate-argument construction

Since the extraction of the subject and the object was carried out successfully, the formation of the triplet as an SAO follows; noun groups of the subject and object are distinguished by labeling without taking into account additional valences.

This structure is used as an intermediate form of data processing. In the future, the primary SAOs are subjected to processing aimed at clarifying the structural elements and linking them together.

3.3 Creation of a tree of relations of elements of structures

The creation of a tree of relationships requires determining the final vertices of the graph, taking into account the presence of identical concepts, as well as identifying implicit relations of generic terms between elements.

For further binding of the primary SAOs, preprocessing is necessary:

- The separation of homogeneous members in the description of structural elements along the border of the composing union "and" provided it is located between the labels of the vertices of the noun group (i.e., "N-SBJ" or "N-OBJ").
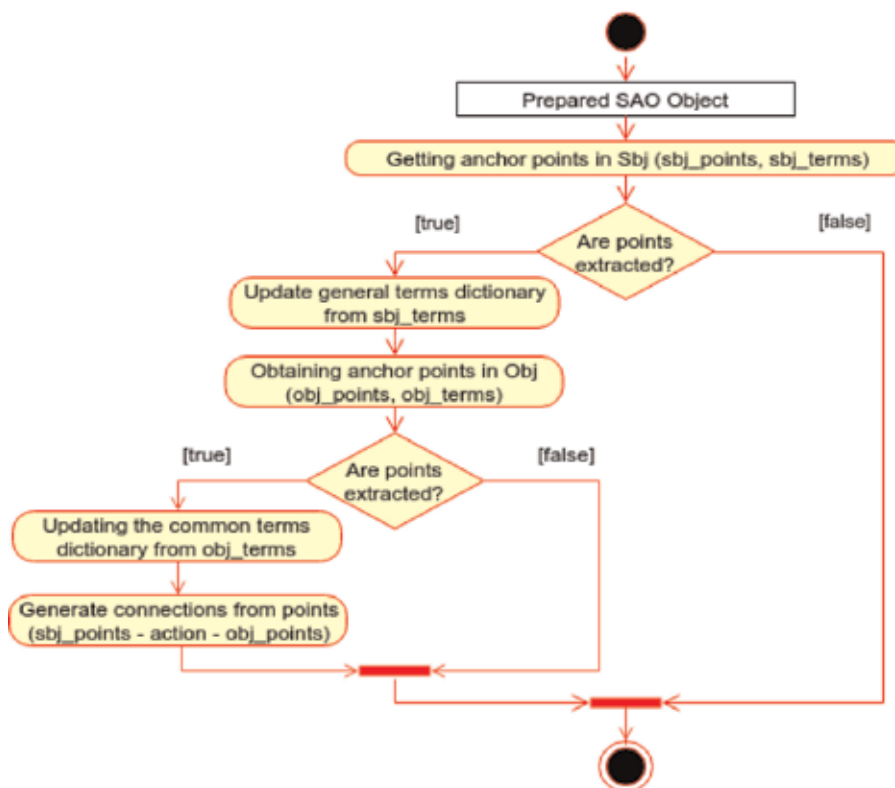

Figure 8.
The general sequence of building links of actants.

- Generation of the normal form of the actants (in the nominative case) and the genitive case (further as the Gent-form).

- After preprocessing, the primary SAO is considered prepared.

When saving an actant, it is necessary to form its normal form (the nominative case), which will be saved as a structural element. For this, the morphological characteristics of the top of the actant are determined, and all previous tokens to the top are put in agreement in the form of the nominative case. For example, "(to) hollow cylinders of stator and rotor" → *"hollow cylinders of rotor and stator"* ("(к) полым цилиндрам статора и ротора" → *"полые цилиндры статора и ротора"* in Russian).

Similar occurs and the declination of the actant is in the form of the genitive case. Gent-form is necessary for finding the parental relationship between the actants, for example, "(with) first output" → *"first output"* ("(с) первым выходом" → *"первого выхода"* in Russian).

Actants are declined using the pymorphy2 tool.

This is followed by linking the prepared SAOs into a single graph. The mechanism consists in the sequential transformation of actants of subjects and objects into a set of anchor points from a common glossary of terms, followed by memorization of the relation (predicate) for the identified points of the subject and object. The overall sequence of building links is represented by the activity diagram in **Figure 8**.

The general graph of relations is stored in the form of a dictionary of terms, a dictionary of relations, and a map of relations containing the parental relations between the actants through predicates (from the dictionary of relations).

Anchor points are formed when analyzing the subject and object of the prepared SAOs. When terms are added (the normal form of an actant's sentence), the parent relationship with another term is searched. The operation is carried out under the condition that the normal form of the actant has a sequence of tokens in the genitive case (marked with an "I-GEN" label in the analysis). An example of a parent relationship search is shown schematically in **Figure 9**.

If the final sequence of tokens in the processed actant is in the genitive case (1), then the word forms in the array of prepared SAO (2) are checked for coincidence. The parent actant index in the terms vocabulary (3) is restored by the bundle identifier or normal form. The processed actant is cleared from the gent-ending, and a new actant is added to the terms vocabulary (5). The "have" parent
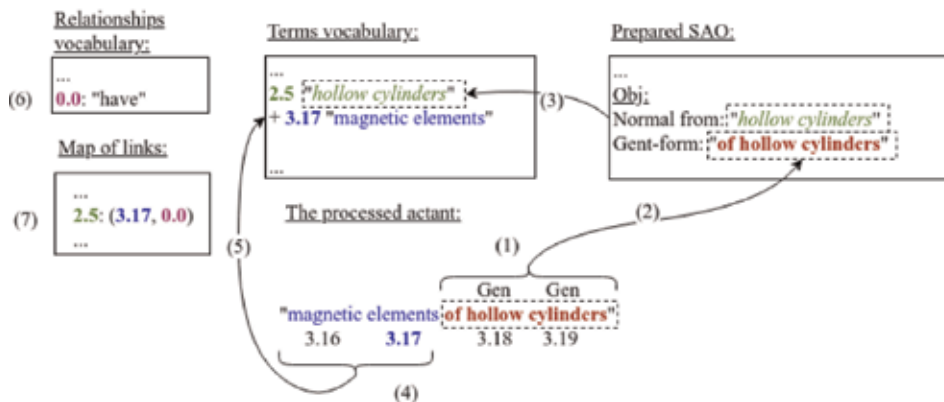

Figure 9.
The search logic of the parent relationship.

relationship is already included by default in the relationship vocabulary (6). Next, the map of links is updated to reflect the new ratio of anchor points (7).

In this case, the indices in the vocabularies are the coordinates of the vertex of the noun group (including the segment number and the token number in the segment).

After receiving the anchor points, consisting of the coordinates and the normal form of the structural elements, the map of connections is filled: the actants of the subject act as parent concepts for the actants of the object connected through the predicate. However, for the semantic class of predicates of relations (e.g., "connect," etc., see **Table 2**), the direction is not so relevant, since it is mutual. The results of processing the prepared SAOs are buffered, and after checking the semantic classes of actants, the graphs of the entire device are updated.

To consider the complete structure of the output, take a small fragment of the text modeling the description of the claims:

"Magnetic gearbox, characterized in that the hollow cylinders are connected with the rotor of slow rotation and with the stator, and the magnetic elements of the hollow cylinders have an angular position" ("Магнитный редуктор, отличающийся тем, что полые цилиндры связаны с ротором медленного вращения и со статором, а магнитные элементы полых цилиндров имеют угловое положение" in Russian).

Primary predicate-argument constructs will be represented by the following SAOs:

Prepared SAO number 1:

Subject	Tokens	hollow cylinders ("полые цилиндры" in Russian)
	Markers	["I-SBJ," "N-SBJ"]
Action		are connected ("связанный" in Russian
Object	Tokens	with the rotor of slow rotation and with the stator ("с ротором медленного вращения и со статором" in Russian)
	Markers	["P-OBJ," "N-OBJ," "I-OBJ," "I-OBJ," "I-OBJ," "P-OBJ," "N-OBJ"]

Prepared SAO number 2:

Subject	Tokens	magnetic elements of the hollow cylinders ("магнитные элементы полых цилиндров" in Russian)
	Markers	["I-SBJ," "N-SBJ," "I-GEN," "I-GEN"]
Action		have ("имеют" in Russian)
Object	Tokens	an angular position ("угловое положение" in Russian)
	Markers	["I-OBJ," "N-OBJ"]

The constructed graph is presented in **Figure 10**.

From example, it can be noted that homogeneous members are distinguished from the object of the bundle SAO No. 1 and the parent relationship of "magnetic elements" ("магнитных элементов" in Russian) and "hollow cylinders" ("полых цилиндров" in Russian) is expressed by the pseudo-relationship "have." In this case, the actant "angular position" ("угловое положение" in Russian) was not added to the output set, since it falls into the forbidden semantic class of words.
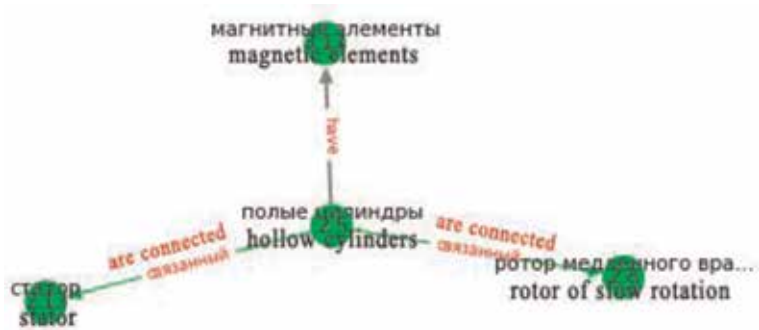
Figure 10.
Constructed graph of structural elements.

4. Construction of domain ontology

The ontology-based patent processing technologies are developing more and more actively. Thus, in [7], ontology extracted concepts and relationships are used to improve patent search. In [29] the information extracted from the claims of the device is stored in an ontological representation and is used for visualization and processing. In [30] patent information in ontology is formalized with reference to technological areas. Thus, the ontological representation provides ample opportunities for the implementation of the description and linking and searching for patent information.

In this paper, at the initial stage, ontology is considered to a greater extent as a storage medium. Inventions and connections between them act as concepts. The designed scheme of the domain ontology is presented in **Figure 11**.
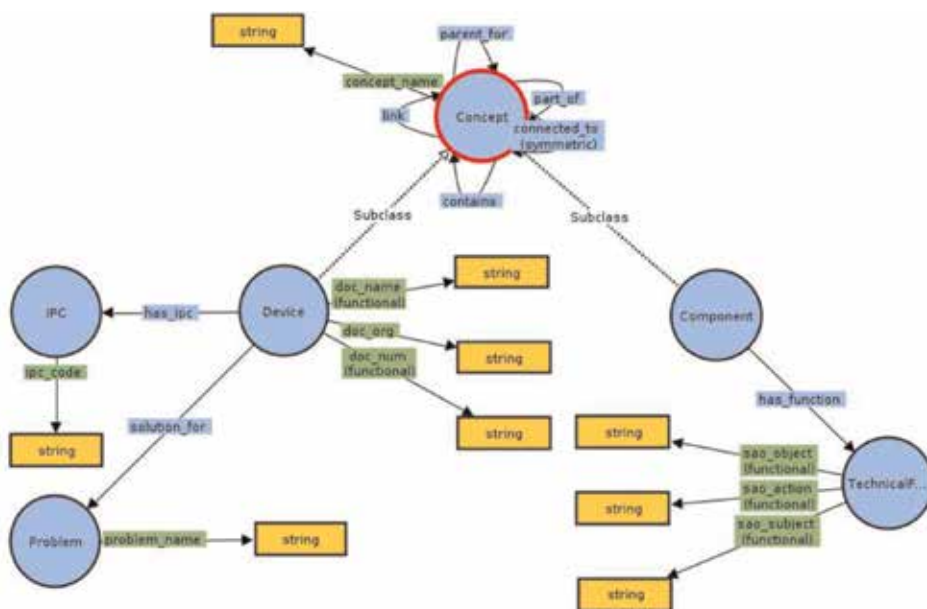
Figure 11.
Ontology scheme for storing patent information.

Figure 12.
The output of all components of a given invention.

The invention (patent document) is assigned with the name of the invention, the patent number, the owner organization, and the IPC codes. Additional concepts of the technical function and the problem solved by the invention are introduced for the subsequent development of the system.

Relationships between components are specified through the following properties:

- connected_to—connection between elements (verbs "set," "connect," etc.)

- contains—indicating the presence of a component (verbs "contain," "have," etc.)

- part_of—an implicit indication of the relationship to the invention

- parent_for—an indication of an implicit parental relationship between the elements (e.g., "hollow stator cylinders")

The ontology graph is described in RDF/XML format. The extracted semantic structures are embedded in the description of ontologies and are uploaded as an OWL file. The resulting data can then be loaded into the RDF storage and make the appropriate SPARQL queries. The data processing processed by the system amounted to 11,200 patent documents.

Further, it is possible to make various requests for information retrieval. An example of a simple request for the conclusion of all the components of a given invention (by the entry of the word "reducer" in the name) is as follows:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX cad: <http://www.vstu.ru/cad/ontology-of-devices#>
SELECT ?device ?dev_name ?component_name
WHERE {
 ?device rdf:type cad:Device .
 ?device cad:concept_name ?dev_name .
 ?device cad:doc_name ?doc_name .
 ?device cad:doc_num ?doc_num .
 ?component cad:part_of ?device .
 ?component cad:concept_name ?component_name
 FILTER(CONTAINS(STR(?dev_name), "редуктор"))
}
```

The result of the query is shown in **Figure 12**.

You can see which inventions use the given component or search for the descendants, etc. At the same time, unlike relational databases, the ontological representation allows you to add descriptive logic and, in general, make more flexible queries to data.

5. System evaluation methodology

The quality of data extraction was assessed for primary SAOs (i.e., excluding post-processing with the separation of homogeneous members and taking into account the semantic classes of actants). For this purpose, manual marking of an independent test set was carried out, including the main points of the claims of 30 patents. The markup was made taking into account the semantic classes of verbs introduced into the system, which characterize the target relations between the components of the technical object (see **Table 2**). In this case, it was allowed to add into the system descriptions (valencies) of previously not encountered target verbs.

The following metrics were used as evaluation metrics: precision, recall, and F1 measure.

Precision is the proportion of correctly extracted engineering implementations (SAO) to the total number of extracted engineering implementations, defined by formula (1):

$$Precision = \frac{|R_{rel}|}{|R_f|} \tag{1}$$

where R_{rel} is the relevant SAO and R_f are all SAOs extracted by the system.

Recall is the ratio of the number of correctly extracted engineering implementations to the total number of relevant structures, defined by formula (2):

$$Recall = \frac{|R_{rel}|}{|S_{rel}|} \tag{2}$$

where R_{rel} is the relevant SAO and S_{rel} are manually selected SAOs.

F1 measure is the harmonic mean of the accuracy and recall; it is determined by formula (3):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

where precision is the extraction accuracy and recall is the extraction completeness.

Parameter	R_f	R_{rel_1}	R_{rel_2}	S_{rel}
Amount	305	198	248	318

Table 7.
Extracted test data set objects.

Parameter	R_f	R_{rel_1}	R_{rel_2}	S_{rel}
Amount	142	118	129	136

Table 8.
Extracted design data set objects.

Since the unit of extraction is a predicate-argument construct, there may be cases of incomplete extraction of one (or both) of its arguments (due to the erroneous identification of the noun group and other reasons). Therefore, two types of evaluation were introduced:

- With a rigorous entry of arguments (there must be a complete occurrence of the selected words)

- With a lax entry of arguments (only the hit of the vertex of the noun group is significant)

Example of mild occurrence of arguments: "the second node is connected to *another* output of the primary winding":
S: the second node
A: is connected
O: to output of the primary winding
In the noun group of the object, the definition of "another" is omitted, and the vertex of the nominal group "output" is present. Therefore, according to the evaluation with a rigorous entry, the structure is considered to be incorrectly recognized. However, by a lax evaluation, the extraction can be considered successful.

Metrics	Test data set		Design data set	
	Lax evaluation	Rigorous evaluation	Lax evaluation	Rigorous evaluation
Precision	0.81	0.65	0.91	0.83
Recall	0.78	0.62	0.95	0.87
F1 measure	0.79	0.63	0.93	0.85

Table 9.
Metric counting results.

Configuration number	Operating system	Device hardware configuration
1	CentOS Linux (7.3.1611) × 64	Intel Xeon CPU E5-2650 v3 @ 2.30 GHz, RAM 132 Gb
2	Ubuntu Linux (18.04) × 64	AMD A8 6410 @ 2GHz, RAM 8 Gb

Table 10.
Hardware configuration.

Program pass	Hardware configuration #1 (cluster)	Hardware configuration #2 (laptop)
Pass 1	2241.3	5614.24
Pass 2	2161.8	5602.21
Pass 3	2151.1	5604.37
Average processing time, sec	2180	5606.94
Average processing time per document, sec	0.581	1.49

Table 11.
The results of the evaluation of the system speed.

An additional evaluation of the quality was also carried out for the design data set of the claims (i.e., on which the system was designed).

A test data set of 30 documents on the basis of manual marking included 318 SAOs. The results of counting the number of extracted test data set objects are presented in **Table 7**.

Designations of headings (according to formulas (1) and (2)):

- R_f—all SAO structures extracted by the system

- R_{rel_1}—relevant SAO structures with a rigorous entry of the entire noun group in the argument

- R_{rel_2}—relevant SAO structures taking into account that the vertex of the noun group is in the argument

- S_{rel}—manually selected SAO structures

The design data set of 14 documents contained 136 recoverable structures. The results of counting the number of extracted SAOs of the design data set are presented in **Table 8**.

The result of the counting of the extraction quality metrics (according to formulas (1)–(3)) for the test and design data set is presented in **Table 9**.

According to the results of the evaluation on the test data set, the proposed method allows to extract data with an accuracy of 63%.

The performance indicator was the average patent processing time (reading from an XML document). The speed of work was determined by the arithmetic average of the time of three runs on a test data set of 3755 documents without marking for each hardware configuration (see **Table 10**).

The results of the run time estimation are presented in **Table 11**.

Patent processing speed is on average less than 1.5 seconds, which is not difficult for batch processing.

6. Conclusion

The general task of this research was extracting information for prior art patent search and new technical solution synthesis based on the analysis of Russian-language patents. The data source was the main claim of the device from the patent text. And the object of extraction was the SAO semantic structures.

Extracting structured data from patents depends on the specifics of the writing of documents. The peculiarities of writing claims of the invention include the excessive length of sentences, the complication of the participle clauses, and characteristic terminology. Practical experience has shown that NLP tools do not always produce a stable result for the patent processing. Thus, they are ineffective in building high-quality data extraction systems with direct application. Therefore, a special approach to the preprocessing of patent texts seems necessary. It is worth noting that in practice they often use a combination of tools to improve the quality of processing.

Due to the sublanguage of patents and the availability of common templates for the formation of documents, the author made an assumption about the sufficiency of morphological and shallow parsing analysis with a number of heuristics to effectively extract structural elements from the description of the claims.

The developed method of extracting SAO structures involves two main stages: the segmentation of the sentences of the claims and the extraction of technical

implementations in the form of predicate-argument structures. Segmentation is based on the derivation of typical structures in the sentence and coordination of dependent words on cases. The extraction of structural elements of technical objects is based on a verb set corresponding to certain semantic classes. After that, the extracted structural elements are linked to a graph, taking into account homogeneous members and parental relations.

The effectiveness of the method was evaluated with an independent test data set of patents with a total number of marked SAO structures of 318 elements. The value of the F1 metric with a rigorous evaluation (full comparison of arguments) and a lax evaluation (the presence of vertexes of noun groups suffices) was 63 and 79%, respectively. The average document processing speed was 1.49 seconds on a laptop with an average configuration.

The precision and recall of Russian-language patent processing exceed the results described in the article [1]. However, not the most advanced level of data extraction of 63% with a rigorous evaluation is due to the inevitable errors of the tools and to the imperfection of the embedded processing logic. It is impossible to take into account all the subtleties of the formation of the sentences especially for a rich Russian language with free word order. The assumption of the finiteness of the segment patterns in the claims is just too strong. However, with a more detailed study, it seems possible for the authors to cover most of the typical variants of writing segments of the claims and bring the system to an industrial level. The results show the prospects of the proposed approach.

The explicit description of many extraction mechanisms (heuristics) hampers the development of the system and the coverage of all possible variations of writing patents. However, the method can still be considered extensible. The inherent stereotyped patent allows not resorting to overly difficult decisions and maintaining a balance between the complexity of the system and its effectiveness.

Further tasks include a more detailed selection of text processing tools, identification of effective methods for statistical extraction of segment patterns, development of heuristics for extracting SAO structures, and improved graph building logic.

The construction of the graph of structural elements of a technical object allows going on to compile the ontology of the subject area and come closer to solving the problems of prior art patent search and information support for new technical solutions synthesis, which is a further direction of the research.

The ontology scheme as a concept includes the structural elements of technical objects and the relationship between them, as well as supporting information on the invention. The initial content of the ontology is based on the processing of 11,200 patent documents for inventions. The existing scheme already allows retrieving useful information about alternatives of structural components and communications between them, for example, searching for all elements of a structure in a given invention or tracking relationships. The results suggest that the proposed approach is promising. A further direction of research is seen by the authors in improving the existing method for extracting data and expanding ontology.

Acknowledgements

Author details

Dmitriy Korobkin*, Sergey Vasiliev, Sergey Fomenkov and S.G. Kolesnikov
Volgograd State Technical University, Russia

*Address all correspondence to: dkorobkin80@mail.ru

IntechOpen

References

[1] Korobkin DM, Fomenkov SA, Kravets AG, Kolesnikov SG. Methods of statistical and semantic patent analysis. In: Kravets A, Shcherbakov M, Kultsova M, Groumpos P. Volgograd State Technical University, et al., editors. CIT and DS 2017: Proceedings. Germany: Springer International Publishing AG; 2017. pp. 48-61. (Ser. Communications in Computer and Information Science; Vol. 754)

[2] Mel'čuk IA. Dependency Syntax: Theory and Practice. NY: SUNY Publ; 1988

[3] Choi S et al. SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. Scientometrics. 2011:863-883. DOI: 10.1007/s11192-011-0420-z

[4] Guo J, Wang X, Li Q, Zhu D, et al. Subject–action–object-based morphology analysis for determining the direction of technological change. Technological Forecasting and Social Change. 2016;**105**:27-40

[5] Wang X et al. Identification of technology development trends based on subject–action–object analysis: The case of dye-sensitized solar cells. Technological Forecasting and Social Change. 2015;**98**:24-46

[6] Yang C et al. SAO semantic information identification for text mining. International Journal of Computational Intelligence Systems. 2017;**10**:593-604. DOI: 10.2991/ijcis.2017.10.1.40

[7] Phan C-P, Nguyen H-Q, Nguyen T-T. Ontology-based heuristic patent search. International Journal of Web Information Systems. 2018. DOI: 10.1108/IJWIS-06-2018-0053

[8] Souili A, Cavallucci D, Rousselot F, Zanni-Merk C. Starting from patents to find inputs to the problem graph model of IDM-TRIZ. Procedia Engineering. 2015;**131**:150-161. DOI: 10.1016/j.proeng.2015.12.365

[9] Supotnitskiy MV. Formula izobreteniya. Vedomosti Nauchnogo tsentra ekspertizy sredstv meditsinskogo primeneniya. 2013;**1**: 41-44

[10] Tomita-parser Developer Guide. Available from: https://tech.yandex.ru/tomita/doc/dg/concept/about-docpage/ [Accessed: 10 May 2019]

[11] Rubaylo AV, Kosenko MYU. Programmnyye sredstva izvlecheniya informatsii iz tekstov na yestestvennom yazyke. Al'manakh sovremennoy nauki i obrazovaniya. 2016;**12**:87-92

[12] Suleymanov RS. Izvlecheniye metadannykh iz polnotekstovykh elektronnykh russkoyazychnykh izdaniy pri pomoshchi Tomita-parsera. Programmnyye produkty i sistemy. 2016;**4**(116):58-62

[13] Koblikov IA, Korobkin DM, Fomenkov SA, Yarovenko VA. Metodika izvlecheniya opisaniy realizuyemykh v patente tekhnicheskikh funktsiy. Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta. 2017; **8**(203):55-59

[14] Link Grammar Parser Homepage. Available from: http://www.abisource.com/projects/linkgrammar [Accessed: 10 May 2019]

[15] MaltParser Homepage. Available from: http://maltparser.org/ [Accessed: 10 May 2019]

[16] UFAL UDPipe Homepage. Available from: http://ufal.mff.cuni.cz/udpipe [Accessed: 10 May 2019]

[17] Korobkin DM, Tyulkina EA, Fomenkov SA, Kolesnikov SG. Sistema izvlecheniya tekhnicheskikh funktsiy iz patentnogo massiva. ITNOU: Informatsionnyye tekhnologii v nauke, obrazovanii i upravlenii. 2017;**2**:24-30

[18] UD Russian SynTagRus Treebank's page. Available from: https://universa ldependencies.org/treebanks/ru_syntag rus/index.html [Accessed: 10 May 2019]

[19] CoNLL-U Format. Available from: https://universaldependencies.org/format.html

[20] Smirnov IV, Salmanov AO, Kuznetsova ES, Haramain IV. Semantiko-sintaksicheskiy analiz yestestvennykh yazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov. Iskusstvennyy intellekt i prinyatiye resheniy, ISA RAN. 2014;**1**: 11-24

[21] Vasiliev SS, Kharitonov AA, Korobkin DM, Fomenkov SA. Extracting descriptions of morphological features of technical objects from Russian-language patents. Modeling, Optimization and Information Technologies. 2018;**4**:6

[22] Kobzareva TYU. V poiskakh sintaksicheskoy struktury: avtomaticheskiy analiz russkogo predlozheniya s oporoy na segmentatsiyu. Moscow: RGGU; 2015

[23] Kharlamov AA, Ermolenko TV, Dorokhina GV. Sravnitel'nyy analiz organizatsii sistem sintaksicheskikh parserov. Inzhenernyy vestnik Dona. 2013;**4**(27):74

[24] Asiryan AK. Morphological tagging tools comparison. In: Intellectual Potential of the XXI Century. 2017. Available from: https://www.sworld. com.ua/konferu7-317/27.pdf [Accessed: 10 May 2019]

[25] Dereza OV, Kayutenko DA, Fenogenova AS. Automatic morphological analysis for Russian: A comparative study. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue". 2016. Available from: http://www.dia log-21.ru/media/3473/dereza.pdf [Accessed: 10 May 2019]

[26] Kuzmenko E. Morphological analysis for Russian: Integration and comparison of taggers. In: Ignatov D et al., editors. Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, Vol 661. Cham: Springer; 2017. pp. 162-171

[27] Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. Analysis of Images, Social Networks and Texts. 2015:320-332. arXiv:1503.07283v1

[28] Fenogenova A. Chanker imennykh grupp russkogo yazyka. Available from: http://web-corpora.net/wsgi/chunker. wsgi/npchunker/npchunker/ [Accessed: 10 May 2019]

[29] Reis SRN, Reis A, Carrabina J, Casanovas P. Contributions to modeling patent claims when representing patent knowledge. Lecture Notes in Computer Science. 2018;**10791**:140-156. DOI: 10.1007/978-3-030-00178-0_9

[30] Ulmschneider K, Glimm B. Semantic exploitation of implicit patent information. In: Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI 2016). Athens, Greece; 2016. DOI:10.1109/ SSCI.2016.7849943

Chapter 5

Density Estimation in Inventory Control Systems under a Discounted Optimality Criterion

Jesús Adolfo Minjárez-Sosa

Abstract

This chapter deals with a class of discrete-time inventory control systems where the demand process $\{D_t\}$ is formed by independent and identically distributed random variables with unknown density. Our objective is to introduce a suitable density estimation method which, combined with optimal control schemes, defines a procedure to construct optimal policies under a discounted optimality criterion.

Keywords: discounted optimality, density estimation, inventory systems, optimal policies, Markov decision processes

AMS 2010 subject classifications: 93E20, 62G07, 90B05

1. Introduction

Inventory systems are one of the most studied sequential decision problems in the fields of operation research and operation management. Its origin lies in the problem of determining how much inventory of a certain product should be kept in existence to meet the demand of buyers, at a cost as low as possible. Specifically, the question is: How much should be ordered, or produced, to satisfy the demand that will be presented during a certain period? Clearly, the behavior of the inventory over time depends on the ordered quantities and the demand of the product in successive periods. Indeed, let I_t and q_t be the inventory level and the order quantity at the beginning of period t, respectively, and D_t be the random demand during period t. Then $\{I_t\}_{t \geq 0}$ is a stochastic process whose evolution in time is given as

$$I_{t+1} = \max\{0, I_t + q_t - D_t\} =: \left(I_t + q_t - D_t\right)^+, \ \ t = 0, 1, \ldots$$

Schematically, this process is illustrated in the following figure.

D_t Demand

Inventory System

Inventory Manager
q_t

$I_{t+1} = (I_t + q_t - D_t)^+$
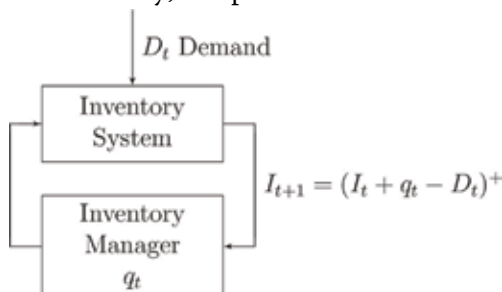
(Standard inventory system)

In this case, the inventory manager (IM) observes the inventory level I_t and then selects the order quantity q_t as a function of I_t. The order quantity process causes costs in the operation of the inventory system. For instance, if the quantity ordered is relatively small, then the items are very likely to be sold out, but there will be unmet demand. In this case the holding cost is reduced, but there is a significant cost due to shortage. Otherwise, if the size of the order is large, there is a risk of having surpluses with a high holding cost. These facts give rise to a stochastic optimization problem, which can be modeled as a Markov decision process (MDP). That is, the inventory system can be analyzed as a stochastic optimal control problem whose objective is to find the optimal ordering policy that minimizes a total expected cost.

The analysis of the control problem associated to inventory systems has been done under several scenarios: discrete-time and continuous-time systems with finite or infinite capacity, inventory systems considering bounded and unbounded one-stage cost, as well as partially observable models, among others (see, e.g., [1–5, 7]). Moreover, such scenarios have their own methods and techniques to solve the corresponding control problem. However, in most cases, it has been assumed that all the components that define the behavior of the inventory system are known to the IM, which, in certain situations, can be too strong and unrealistic. Hence it is necessary to implement schemes that allow learning or collecting information about the unknown components during the evolution of the system to choose a decision with as much information as possible.

In this chapter we study a class of inventory control systems where the density of the demand is unknown by the IM. In this sense, our objective is to propose a procedure that combines density estimation methods and control schemes to construct optimal policies under a total expected discounted cost criterion. The estimation and control procedure is illustrated in the following figure:

$$I_{t+1} = (I_t + q_t - D_t)^+$$
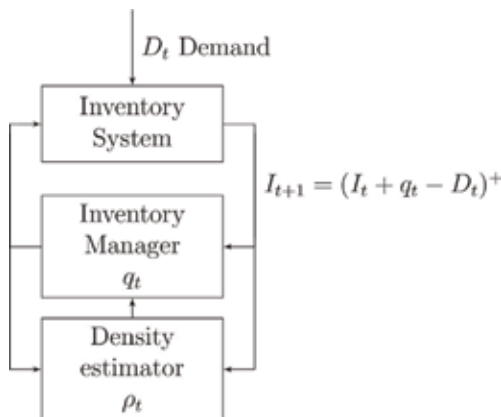
(Estimation and control procedure)

In this case, unlike the standard inventory system, before choosing the order quantity q_t, the IM implements a density estimation method to get an estimate ρ_t, and, possibly, combines this with the history of the system $h_t = (I_0, q_0, D_0, ..., I_{t-1}, q_{t-1}, D_{t-1}, I_t)$ to select $q_t = q_t(h_t, \rho_t)$. Specifically, the density of the demand is estimated by the projection of an arbitrary estimator on an appropriate set, and its convergence is stated with respect to a norm which depends on the components of the inventory control model.

In general terms, our approach consists in to show that the inventory system can be studied under the weighted-norm approach, widely studied by several authors in

the field of Markov decision processes (see, e.g., [11] and references therein) and in adaptive control (see, e.g. [9, 12–14]). That is, we prove the existence of a weighted function *W* which imposes a growth condition on the cost functions. Then, applying the dynamic programming algorithm, the density estimation method is adapted to such a condition to define an estimation and control procedure for the construction of optimal policies.

The chapter is organized as follows. In Section 2 we describe the inventory model and define the corresponding optimal control problem. In Section 3 we introduce the dynamic programming approach under the true density. Next, in Section 4 we present the density estimation method which will be used to state, in Section 5, an estimation and control procedure for the construction of optimal policies. The proofs of the main results are given in Section 6. Finally, in Section 7, we present some concluding remarks.

2. The inventory model

We consider an inventory system evolving according to the difference equation

$$I_{t+1} = \left(I_t + q_t - D_t\right)^+, \quad t = 0, 1, ..., \tag{1}$$

where I_t and q_t are the inventory level and the order quantity at the beginning of period t, taking values in $\mathbb{I} := [0, \infty)$ and $\mathbb{Q} := [0, \infty)$, respectively, and D_t represents the random demand during period t. We assume that $\{D_t\}$ is an observable sequence of nonnegative independent and identically distributed (i.i.d.) random variables with a common density $\rho \in L_1[0, \infty)$ which is unknown by the inventory manager. In addition, we assume finite expectation

$$\overline{D} := E(D_t) < \infty. \tag{2}$$

Moreover, there exists a measurable function $\overline{\rho} \in L_1[0, \infty)$ such that

$$\rho(s) \leq \overline{\rho}(s) \tag{3}$$

almost everywhere with respect to the Lebesgue measure. In addition

$$\int_0^\infty s^2 \overline{\rho}(s) ds < \infty. \tag{4}$$

For example, if $\overline{\rho}(s) := K \min\{1, 1/s^{1+r}\}, s \in [0, \infty)$, for some positive constants K and r, then there are plenty of densities that satisfy (3)–(4).

The one-stage cost function is defined as

$$\tilde{c}(I, q, D) = cq + h(I + q - D)^+ + b(D - I - q)^+, \quad (I, q) \in \mathbb{I} \times \mathbb{Q}, \tag{5}$$

where $h, c,$ and b are, respectively, the holding cost per unit, the ordering cost per unit, and the shortage cost per unit, satisfying $b > c$.

The order quantities applied by the IM are selected according to rules known as ordering control policies defined as follows. Let $\mathcal{H}_t$ be the space of histories of the inventory system up to time t. That is, a typical element of $\mathcal{H}_t$ is written as

$$h_t = \left(I_0, q_0, D_0, ..., I_{t-1}, q_{t-1}, D_{t-1}, I_t\right).$$

An ordering policy (or simply a policy) $\gamma = \{\gamma_t\}$ is a sequence of measurable functions $\gamma_t : \mathcal{H}_t \to \mathbb{Q}$, such that $\gamma_t(h_t) = q_t, t \geq 0$. We denote by Γ the set of all policies. A feedback policy or Markov policy is a sequence $\gamma = \{g_t\}$ of functions $g_t : \mathbb{I} \to \mathbb{Q}$, such that $g_t(I_t) = q_t$. A feedback policy $\gamma = \{g_t\}$ is stationary if there exists a function $g : \mathbb{I} \to \mathbb{Q}$ such that $g_t = g$ for all $t \geq 0$.

When using a policy $\gamma \in \Gamma$, given the initial inventory level $I_0 = I$, we define the total expected discounted cost as

$$V(\gamma, I) := E\left[\sum_{t=0}^{\infty} \alpha^t \tilde{c}\left(I_t, q_t, D_t\right)\right], \tag{6}$$

where $\alpha \in (0, 1)$ is the so-called discount factor. The inventory control problem is then to find an optimal feedback policy γ^* such that $V(\gamma^*, I) = V^*(I)$ for all $I \in \mathbb{I}$, where

$$V^*(I) := \inf_{\gamma \in \Gamma} V(\gamma, I), \quad I \in \mathbb{I}, \tag{7}$$

is the optimal discounted cost, which we call value function.

We define the mean one-stage cost as

$$c(I, q) = cq + hE(I + q - D)^+ + bE(D - I - q)^+$$
$$= cq + h \int_0^{I+q} (I + q - s)^+ \rho(s) ds + b \int_{I+q}^{\infty} (s - I - q)^+ \rho(s) ds, \quad (I, q) \in \mathbb{I} \times \mathbb{Q}. \tag{8}$$

Then, by using properties of conditional expectation, we can rewrite the total expected discounted cost (6) as

$$V(\gamma, I) = E_I^{\gamma}\left[\sum_{t=0}^{\infty} \alpha^t c\left(I_t, q_t\right)\right], \tag{9}$$

where E_I^{γ} denotes the expectation operator with respect to the probability P_I^{γ} induced by the policy γ, given the initial inventory level $I_0 = I$ (see, e.g., [8, 10]).

The sequence of events in our model is as follows. Since the density ρ is unknown, the one-stage cost (8) is also unknown by the IM. Then if at stage t the inventory level is $I_t = I \in \mathbb{I}$, the IM implements a suitable density estimation method to get an estimate ρ_t of ρ. Next, he/she combines this with the history of the system to select an order quantity $q_t = q = \gamma_t^{\rho_t}(h_t) \in \mathbb{Q}$. Then a cost $c(I, q)$ is incurred, and the system moves to a new inventory level $I_{t+1} = I' \in \mathbb{I}$ according to the transition law

$$Q(B|I, q) := \text{Prob}\left[I_{t+1} \in B | I_t = I, q_t = q\right]$$
$$= \int_0^{\infty} 1_B\left((I + q - s)^+\right) \rho(s) ds \tag{10}$$

where $1_B(.)$ denotes the indicator function of the set $B \in \mathcal{B}(\mathbb{I})$, and $\mathcal{B}(\mathbb{I})$ is the Borel $\sigma-$algebra on $\mathbb{I}$. Once the transition to the inventory level I' occurs, the

process is repeated. Furthermore, the costs are accumulated according to the discounted cost criterion (9).

3. Dynamic programming equation under the true density ρ

The study of the inventory control problem will be done by means of the well-known dynamic programming (DP) approach, which we now introduce in terms of the unknown density ρ. In order to establish precisely the ideas, we first present some preliminary and useful facts.

The set of order quantities in which we can find the optimal ordering policy should be $\mathbb{Q}^* = [0, Q^*] \subset \mathbb{Q}$,

where

$$Q^* = \frac{b\overline{D}}{c(1-\alpha)}.$$

Thus, we can restrict the range of q so that $q \in \mathbb{Q}^*$. Specifically we have the following result.

Lemma 3.1 *Let $\gamma^0 \in \Gamma$ be the policy defined as $\gamma^0 = \{0, 0, ...\}$, and let $\overline{\gamma} = \{\overline{\gamma}_t\}$ be a policy such that $\overline{\gamma}_k(h_k) = \overline{q}_k > Q^*$, for at least a $k = 0, 1, ...$ Then*

$$V(\gamma^0, I) \leq V(\overline{\gamma}, I), \quad I \in \mathbb{I}. \tag{11}$$

That is, γ^0 is a better solution than $\overline{\gamma}$.

Proof. Let I_t^0, $t = 0, 1, ...,$ be the inventory levels generated by the application of γ^0, and $(\overline{I}_t, \overline{q}_t)$ be the sequence of inventory levels and order quantities generated by $\overline{\gamma}$, where $I_0^0 = \overline{I}_0 = I$, $I_{t+1}^0 = (I_t^0 - D_t)^+$, and $\overline{I}_{t+1} = (\overline{I}_t + \overline{q}_t - D_t)^+$, $t \geq 0$. Without loss of generality, we suppose that for a $\overline{q} > Q^*$ we have $\overline{q}_0 = \overline{q}$. Note that $I_t^0 \leq \overline{I}_t$, for all $t \geq 0$. Then observing that $c\overline{q} > b\overline{D}/(1-\alpha)$,

$$V(\gamma^0, I) = E\left[\sum_{t=0}^{\infty} \alpha^t c(I_t^0, 0, D_t)\right] = E\left[\sum_{t=0}^{\infty} \alpha^t \left(h(I_t^0 - D_t)^+ + b(D_t - I_t^0)^+\right)\right]$$

$$\leq E\sum_{t=0}^{\infty} \alpha^t h(\overline{I}_t - D_t)^+ + b\sum_{t=0}^{\infty} \alpha^t E(D_t)$$

$$\leq E\left[\sum_{t=0}^{\infty} \alpha^t \left(h(\overline{I}_t + \overline{q}_t - D_t)^+ + b(D_t - \overline{I}_t - \overline{q}_t)^+ + \frac{b\overline{D}}{1-\alpha}\right)\right]$$

$$\leq E\left[\sum_{t=0}^{\infty} \alpha^t \left(h(\overline{I}_t + \overline{q}_t - D_t)^+ + b(D_t - \overline{I}_t - \overline{q}_t)^+ + c\overline{q}\right)\right]$$

$$\leq E\left[\sum_{t=0}^{\infty} \alpha^t \left(c\overline{q}_t + h(\overline{I}_t + \overline{q}_t - D_t)^+ + b(D_t - \overline{I}_t - \overline{q}_t)^+\right)\right]$$

$$= V(\overline{\gamma}, I), \quad I \in \mathbb{I}. \blacksquare$$

Remark 3.2 *Observe that for $(I, q) \in \mathbb{I} \times \mathbb{Q}^*$ we have*

$$c(I, q) = cq + L(I + q),$$

where, by writing $y = I + q$,

$$L(y) := hE(y - D)^+ + bE(D - y)^+.$$

In addition, observe that for any fixed $s \in [0, \infty)$, the functions $y \to (y - s)^+$ and $y \to (s - y)^+$ are convex, which implies that $L(y)$ is convex. Moreover

$$\lim_{y \to \infty} L(y) = \infty.$$

The following lemma provides a growth property of the one-stage cost function (8).

Lemma 3.3 *There exist a number β and a function $W : \mathbb{I} \to [1, \infty)$ such that $0 < \alpha\beta < 1$,*

$$\sup_{(I,q,s) \in \mathbb{I} \times \mathbb{Q}^* \times [0, \infty)} \frac{W\big((I + q - s)^+\big)}{W(I)} := \varphi < \infty, \tag{12}$$

and for all $(I, q) \in \mathbb{I} \times \mathbb{Q}^*$

$$c(I, q) \leq W(I). \tag{13}$$

In addition, for any density μ on $[0, \infty)$ such that $\int_0^\infty s\mu(s) < \infty$,

$$\int_0^\infty W\big((I + q - s)^+\big)\mu(s)ds \leq \beta W(I), \quad (I, q) \in \mathbb{I} \times \mathbb{Q}^*. \tag{14}$$

The proof of Lemma 3.3 is given in Section 6.

We denote by $\mathbb{B}_W$ the normed linear space of all measurable functions $u : \mathbb{I} \to \mathfrak{R}$ with finite weighted-norm (W−norm) $\| \cdot \|_W$ defined as

$$\|u\|_W := \sup_{I \in \mathbb{I}} \frac{|u(I)|}{W(I)}. \tag{15}$$

Essentially, Lemma 3.3 proves that the inventory system (1) falls within of the weighted-norm approach used to study general Markov decision processes (see, e.g., [11]). Hence, we can formulate, on the space $\mathbb{B}_W$, important results as existence of solutions of the DP-equation, convergence of the value iteration algorithm, as well as existence of optimal policies, in the context of the inventory system (1). Indeed, let

$$V^{(n)}(\gamma, I) = E_I^\gamma \left[\sum_{t=0}^{n-1} \alpha^t c\big(I_t, q_t\big) \right]$$

be the n-stage discounted cost under the policy $\gamma \in \Gamma$ and the initial inventory level $I \in \mathbb{I}$, and

$$V^{(n)}(I) = \inf_{\gamma \in \Gamma} V^{(n)}(\gamma, I); \quad V^{(0)}(I) = 0, \quad I \in \mathbb{I}$$

the corresponding value function. Then, for all $n \geq 0$ and $I \in \mathbb{I}$, (see, e.g., [6, 10, 11]),

$$V^{(n)}(I) = \min_{q \in \mathbb{Q}^*} \left\{ c(I,q) + \alpha \int_0^\infty V^{(n-1)}\big((I+q-s)^+\big)\rho(s)ds \right\} \qquad (16)$$

Moreover, from [11, Theorem 8.3.6], by making the appropriate changes, we have the following result.

Theorem 3.4 (Dynamic programming) *(a) The functions $V^{(n)}$ and V^* belong to $\mathbb{B}_W$. Moreover*

$$V^{(n)}(I) \le \frac{W(I)}{1-\alpha\beta}, V^*(I) \le \frac{W(I)}{1-\alpha\beta}, I \in \mathbb{I}. \qquad (17)$$

(b) As $n \to \infty$, $\left\| V^{(n)} - V^* \right\|_W \to 0$.
(c) V^* is convex.
(d) V^* satisfies the dynamic programming equation:

$$V^*(I) = \min_{q \in \mathbb{Q}^*} \left\{ c(I,q) + \alpha \int_0^\infty V^*\big((I+q-s)^+\big)\rho(s)ds \right\}$$

$$= \min_{I \le y \le Q^*+I} \left\{ cy + L(y) + \alpha \int_0^\infty V^*\big((y-s)^+\big)\rho(s)ds \right\} - cI, I \in \mathbb{I}. \qquad (18)$$

(e) There exists a function $g^* : \mathbb{I} \to \mathbb{Q}$ such that $g^*(I) \in \mathbb{Q}^*$ and, for each $I \in \mathbb{I}$,

$$V^*(I) = c(I,g^*(I)) + \alpha \int_0^\infty V^*\big((I+g^*(I)-s)^+\big)\rho(s)ds, I \in \mathbb{I}.$$

Moreover, $\gamma^* = \{g^*\}$ is an optimal control policy.

4. Density estimation

As the density ρ is unknown, the results in Theorem 3.4 are not applicable, and therefore they are not accessible to the IM. In this section we introduce a suitable density estimation method with which we can obtain an estimated DP-equation. This will allow us to define a scheme for the construction of optimal policies. To this end, let $D_0, D_1, ..., D_t, ...$ be independent realizations of the demand whose density is ρ.

Theorem 4.1 *There exists an estimator $\rho_t(s) := \rho_t(s; D_0, D_1, ..., D_{t-1})$, $s \in 0, \infty)$, of ρ, such that (see (2) and (3)):*
  D.1. $\rho_t \in L_1[0, \infty)$ *is a density.*
  D.2. $\rho_t \le \overline{\rho}(\cdot)$ *a.e. with respect to the Lebesgue measure.*
  D.3. $\int_0^\infty s\rho_t(s)ds \le \overline{D}$.
  D.4. $E \int_0^\infty |\rho_t - \rho(s)|ds \to 0$, *as $t \to \infty$.*
  D.5. $E\|\rho_t - \rho\| \to 0$, *as $t \to \infty$, where*

$$\|\mu\| := \sup_{(I,q) \in \mathbb{I} \times \mathbb{Q}^*} \frac{1}{W(I)} \int_0^\infty W\big((I+q-s)^+\big)\mu(s)ds \qquad (19)$$

for measurable functions μ on $[0, \infty)$.

It is worth noting that for any density μ on $[0, \infty)$ satisfying (14), the norm $\|\mu\|$ is finite. The remainder of the section is devoted to prove Theorem 4.1.

We define the set $\mathcal{D} \subset L_1([0, \infty))$ as:

$$\mathcal{D} := \left\{ \mu : \mu \text{ is a density}, \quad \int_0^\infty s\mu(s)ds \leq \overline{D}, \quad \mu(s) \leq \overline{\rho}(s) \text{ a.s.} \right\}.$$

Observe that $\rho \in \mathcal{D}$.

Lemma 4.2 *The set $\mathcal{D}$ is closed and convex in $L_1([0, \infty))$.*

Proof. The convexity of $\mathcal{D}$ follows directly. To prove that $\mathcal{D}$ is closed, let $\mu_t \in \mathcal{D}$ be a sequence in $\mathcal{D}$ such that $\mu_t \xrightarrow{L_1} \mu \in L_1([0, \infty))$. First, we prove

$$\mu(s) \leq \overline{\rho}(s) \quad a.e. \tag{20}$$

We assume that there is $A \subset [0, \infty)$ with $m(A) > 0$ such that $\mu(s) > \overline{\rho}(s), s \in A,$ m being the Lebesgue measure on $\mathfrak{R}$. Then, for some $\varepsilon > 0$ and $A' \subset A$ with $m(A') > 0$,

$$\mu(s) > \overline{\rho}(s) + \varepsilon, s \in A'. \tag{21}$$

Now, since $\mu_t \in \mathcal{D}, t \geq 0,$ there exists $B_t \subset 0, \infty)$ with $m(B_t) = 0,$ such that

$$\mu_t(s) \leq \overline{\rho}(s), \quad s \in [0, \infty)\backslash B_t, t \geq 0. \tag{22}$$

Combining (21) and (22) we have

$$|\mu_t(s) - \mu(s)| \geq \varepsilon, \quad s \in A' \cap ([0, \infty)\backslash B_t), \quad t \geq 0.$$

Using the fact that $m(A' \cap ([0, \infty)\backslash B_t) = m(A') > 0$, we obtain that μ_t does not converge to μ in measure, which is a contradiction to the convergence in L_1. Therefore $\mu(s) \leq \overline{\rho}(s)$ a.e.

On the other hand, applying Holder's inequality and using the fact that $\overline{\rho} \in L_1[0, \infty)$, from (20),

$$\left| 1 - \int_0^\infty \mu(s)ds \right| = \left| \int_0^\infty \mu_t(s)ds - \int_0^\infty \mu(s)ds \right| = \int_0^\infty |\mu_t(s) - \mu(s)|^{\frac{1}{2}}|\mu_t(s) - \mu(s)|^{\frac{1}{2}}ds$$

$$\leq \left(\int_0^\infty 2\overline{\rho}(s) \right)^{1/2} \left(\int_0^\infty |\mu_t(s) - \mu(s)| \right)^{1/2} \to 0 \text{ as } t \to \infty, \tag{23}$$

which implies $\int_0^\infty \mu(s)ds = 1$. Now, as $\mu \geq 0$ *a.e.,* we have that μ is a density. Similarly, from (4),

$$\int_0^\infty s|\mu_t(s) - \mu(s)|ds = \int_0^\infty s|\mu_t(s) - \mu(s)|^{\frac{1}{2}}|\mu_t(s) - \mu(s)|^{\frac{1}{2}}ds$$

$$\leq \left(\int_0^\infty s^2 2\overline{\rho}(s)ds \right)^{1/2} \left(\int_0^\infty |\mu_t(s) - \mu(s)|ds \right)^{1/2} \tag{24}$$

$$\leq 2^{\frac{1}{2}}M' \left(\int_0^\infty |\mu_t(s) - \mu(s)|ds \right)^{1/2},$$

for some constant $M' < \infty$. Letting $t \to \infty$ we obtain

$$\int_0^\infty s\mu_t(s)ds \to \int_0^\infty s\mu(s)ds$$

which, in turn, implies that

$$\int_0^\infty s\mu(s)ds \le \overline{D}.$$

This proves that $\mathcal{D}$ is closed.∎

Let $\hat{\rho}_t(s) := \hat{\rho}_t(s; D_0, D_1, ..., D_t)$, $s \in [0, \infty)$, be an arbitrary estimator of ρ such that

$$E\|\rho - \hat{\rho}_t\|_{L_1} = E\int_0^\infty |\rho(s) - \hat{\rho}_t(s)| \to 0 \quad \text{as } t \to \infty. \tag{25}$$

Lemma 4.2 ensures the existence of the estimator ρ_t which is defined by the projection of $\hat{\rho}_t$ on the set of densities $\mathcal{D}$. That is, the density $\rho_t \in \mathcal{D}$, expressed as

$$\rho_t := \arg\min_{\sigma \in \mathcal{D}} \|\sigma - \hat{\rho}_t\|_{L_1},$$

is the "best approximation" of the estimator $\hat{\rho}_t$ on the set $\mathcal{D}$, that is,

$$\|\rho_t - \hat{\rho}_t\|_{L_1} = \inf_{\mu \in \mathcal{D}} \|\mu - \hat{\rho}_t\|_{L_1}. \tag{26}$$

Now observe that ρ_t satisfies the properties D.1, D.2, and D.3. Hence, Theorem 4.1 will be proved if we show that ρ_t satisfies D.4 and D.5. To this end, since $\rho \in \mathcal{D}$, from (26) observe that

$$\|\rho_t - \rho\|_{L_1} \le \|\rho_t - \hat{\rho}_t\|_{L_1} + \|\hat{\rho}_t - \rho\|_{L_1} \le 2\|\hat{\rho}_t - \rho\|_{L_1}, \; t \ge 0,$$

which implies that, from (25),

$$E\int_0^\infty |\rho(s) - \rho_t(s)|ds \le 2E\|\hat{\rho}_t - \rho\|_{L_1} \to 0, \text{ as } t \to \infty. \tag{27}$$

That is, ρ_t satisfies Property D.4. In fact, since $\int_0^\infty |\rho(s) - \rho_t(s)|ds \le 2$ a.s., from (27) it is easy to see that

$$E\left(\int_0^\infty |\rho(s) - \rho_t(s)|ds\right)^q \to 0, \text{ as } t \to \infty, \text{ for any } q > 0. \tag{28}$$

Now, to obtain property D.5, observe that from (12)

$$\|\rho_t - \rho\| = \sup_{(I,q) \in \mathbb{I} \times \mathbb{Q}^*} \frac{1}{W(I)} \int_0^\infty W((I + q - s)^+)|\rho(s) - \rho_t(s)|ds = \varphi \int_0^\infty |\rho(s) - \rho_t(s)|ds. \tag{29}$$

Therefore, property D.4 yields

$$E\|\rho_t - \rho\| \to 0, \quad \text{as } t \to \infty, \tag{30}$$

which proves the property D.5.

5. Estimation and control

Having defined the estimator ρ_t, we will now introduce an estimate dynamic programming procedure with which we can construct optimal policies for the inventory systems.

Observe that for each $t \geq 0$, from (14),

$$\int_0^\infty W\big((I + q - s)^+\big)\rho_t(s)ds \leq \beta W(I), \quad (I, q) \in \mathbb{I} \times \mathbb{Q}^*. \tag{31}$$

Now, we define the estimate one-stage cost function:

$$c_t(I, q) = cq + h \int_0^{I+q} (I + q - s)^+ \rho_t(s)ds + b \int_{I+q}^\infty (s - I - q)^+ \rho_t(s)ds \tag{32}$$

$$= cq + L_t(I + q), \quad (I, q) \in \mathbb{I} \times \mathbb{Q}^*,$$

where (see Remark 3.2) for $y = I + q$,

$$L_t(y) := h \int_0^y (y - s)^+ \rho_t(s)ds + b \int_y^\infty (s - y)^+ \rho_t(s)ds.$$

In addition, observe that for each $t \geq 0$, $L_t(y)$ is convex and

$$\lim_{y \to \infty} L_t(y) = \infty. \tag{33}$$

We define the sequence of functions $\{V_t\}$ as $V_0 \equiv 0$, and for $t \geq 1$

$$V_t(I) = \min_{q \in \mathbb{Q}^*} \left\{ c_t(I, q) + \alpha \int_0^\infty V_{t-1}\big((I + q - s)^+\big)\rho_t(s)ds \right\} \tag{34}$$

$$= \min_{I \leq y \leq Q^* + I} \left\{ cy + L_t(y) + \alpha \int_0^\infty V_{t-1}\big((y - s)^+\big)\rho_t(s)ds \right\} - cI, \quad I \in \mathbb{I}.$$

We can state our main results as follows:

Theorem 5.1 *(a) For $t \geq 0$ and $I \in \mathbb{I}$,*

$$V_t(I) \leq \frac{W(I)}{1 - \alpha\beta}. \tag{35}$$

Therefore, $V_t \in \mathbb{B}_W$.

(b) As $t \to \infty$, $E\left[\sup_{(I,q) \in \mathbb{I} \times \mathbb{Q}^*} \dfrac{|c_t(I,q) - c(I,q)|}{W(I)} \right] \to 0.$

(c) As $t \to \infty$, $E\|V_t - V^*\|_W \to 0$.

(d) For each $t \geq 0$, there exists $K_t \geq 0$ such that the selector $g_t : \mathbb{I} \to \mathbb{Q}$ defined as

$$q_t^* = g_t(I) := \begin{cases} K_t - I & if \quad 0 \leq I \leq K_t \\ 0 & if \quad\quad I > K_t \end{cases}$$

attains the minimum in (34).

Remark 5.2 *From [10, Proposition D.7], for each $I \in \mathbb{I}$, there is an accumulation point $g_\infty(I) \in \mathbb{Q}^*$ of the sequence $\{g_t(I)\}$. Hence, there exists a constant K^* such that*

$$g_\infty(I) := \begin{cases} K^* - I & if \quad 0 \leq I \leq K^* \\ 0 & if \quad\quad I > K^* \end{cases} \tag{36}$$

Theorem 5.3 *Let g_∞ be the selector defined in (36). Then the stationary policy $\gamma^* := \{g_\infty\}$ is an optimal base stock policy for the inventory problem.*

6. Proofs

6.1 Proof of Lemma 3.3

Note that, for each $(I,q) \in \mathbb{I} \times \mathbb{Q}^*$,

$$\begin{aligned} c(I,q) &\leq cQ^* + h(I + Q^*) + b\overline{D} \\ &\leq (c+h)Q^* + hI + b\overline{D} \leq MG(I), \end{aligned} \tag{37}$$

where $M := \max\{(c+h)Q^* + b\overline{D}, h\}$ and $G(I) = I + 1$. Moreover, for every density function μ on $[0, \infty)$ and $(I,q) \in \mathbb{I} \times \mathbb{Q}^*$,

$$\int_0^\infty G\big((I+q-s)^+\big)\mu(s)ds \leq G(I) + Q^*. \tag{38}$$

On the other hand, we define the sequence of functions $\{w_t\}$, $w_t : \mathbb{I} \to \Re$, as

$$w_0(I) := 1 + MG(I) \tag{39}$$

and for $t \geq 1$ and any density function μ on $[0, \infty)$

$$w_t(I) := \sup_{q \in \mathbb{Q}^*} \int_0^\infty w_{t-1}\big((I+q-s)^+\big)\mu(s)ds.$$

Observe that, for each $I \in \mathbb{I}$,

$$w_1(I) = \sup_{q \in \mathbb{Q}^*} \int_0^\infty \big[1 + MG\big((I+q-s)^+\big)\big]\mu(s)ds$$

$$\leq 1 + MG(I) + MQ^*.$$

Thus,

$$w_2(I) = \sup_{q \in \mathbb{Q}^*} \int_0^\infty \left[1 + MG\big((I + q - s)^+ \big) + MQ^* \right] \mu(s) ds$$

$$\leq 1 + MG(I) + MQ^* + MQ^*, \quad I \in \mathbb{I}.$$

In general, it is easy to see that for each $I \in \mathbb{I}$,

$$w_t(I) \leq MG(I) + 1 + \sum_{j=0}^{t-1} MQ^* = MG(I) + 1 + MQ^* t. \tag{40}$$

Let $\alpha_0 \in (\alpha, 1)$ be arbitrary, and define

$$W(I) := \sum_{t=0}^\infty \alpha_0^t w_t(I). \tag{41}$$

Then, from (40),

$$W(I) \leq \sum_{t=0}^\infty \alpha_0^t [MG(I) + 1 + MQ^* t]$$

$$= \sum_{t=0}^\infty \alpha_0^t (MG(I) + 1) + MQ^* \sum_{t=0}^\infty t\alpha_0^t \leq \frac{MG(I) + 1}{1 - \alpha_0} + \frac{MQ^* \alpha_0}{(1 - \alpha_0)^2}. \tag{42}$$

Therefore, $W(I) < \infty$ for each $I \in \mathbb{I}$, and since $w_0 > 1$, from (41),

$$W(I) > 1. \tag{43}$$

Furthermore, using (42) and the fact that $W(\cdot) \geq w_0(\cdot)$, a straightforward calculation shows that

$$\varphi := \sup_{(I,q,s) \in \mathbb{I} \times \mathbb{Q}^* \times 0, \infty)} \frac{W\big((I + q - s)^+ \big)}{W(I)} < \infty. \tag{44}$$

Now, from (37) and (39), $c(I, q) \leq w_0(I)$, which yields, for all $(I, q) \in \mathbb{I} \times \mathbb{Q}^*$,

$$c(I, q) \leq W(I). \tag{45}$$

In addition, for every density function μ on $[0, \infty)$ and $(I, q) \in \mathbb{I} \times \mathbb{Q}^*$,

$$\int_0^\infty W\big((I + q - s)^+ \big) \mu(s) ds = \int_0^\infty \sum_{t=0}^\infty \alpha_0^t w_t \big((I + q - s)^+ \big) \mu(s) ds$$

$$= \sum_{t=0}^\infty \alpha_0^t \int_0^\infty w_t \big((I + q - s)^+ \big) \mu(s) ds$$

$$\leq \sum_{t=0}^\infty \alpha_0^t w_{t+1}(I) = \alpha_0^{-1} \left[\sum_{t=0}^\infty \alpha_0^t w_t(I) - w_0(I) \right]$$

$$= \alpha_0^{-1} [W(I) - w_0(I)] \leq \alpha_0^{-1} W(I).$$

Therefore, defining $\beta := \alpha_0^{-1}$, we have that $0 < \alpha\beta < 1$, and

$$\int_0^\infty W\big((I + q - s)^+\big)\mu(s)ds \le \beta W(I), \;\; (I,q) \in \mathbb{I} \times \mathbb{Q}^*,$$

which, together with (43), (44), and (45), proves Lemma 3.3.∎

6.2 Proof of Theorem 5.1

(a) Since $\int_0^\infty s\rho_t(s)ds \le \overline{D}$, from (32) (see (37)) $c_t(I,q) \le MG(I)$ for each $t \ge 0$, $(I,q) \in \mathbb{I} \times \mathbb{Q}^*$. Hence, it is easy to see that $c_t(I,q) \le W(I)$ for each $(I,q) \in \mathbb{I} \times \mathbb{Q}^*$ (see (45)). Then we have $V_1(I) \le W(I)$, and from (31), and by applying induction arguments, we get

$$V_t(I) \le \frac{W(I)}{1 - \alpha\beta}, \;\; t \ge 0, \;\; I \in \mathbb{I}. \tag{46}$$

(b) Observe that from (39), for each $I \in \mathbb{I}$,

$$W(I) \ge w_0(I) = 1 + MG(I),$$

which implies that (see (43))

$$\frac{MG(I)}{W(I)} \le 1 - \frac{1}{W(I)} < \infty. \tag{47}$$

In addition, from (37),

$$h(I + Q^*) \le MG(I). \tag{48}$$

On the other hand, similarly as (24), from (4), it is easy to see that

$$\int_0^\infty s|\rho_t(s) - \rho(s)|ds \le 2^{\frac{1}{2}}M'\left(\int_0^\infty |\rho_t(s) - \rho(s)|ds\right)^{1/2}, \tag{49}$$

for some constant $M' < \infty$. Hence, combining (47)–(49), from the definition of $c_t(I,q)$ and $c(I,q)$, we have

$$\frac{|c_t(I,q) - c(I,q)|}{W(I)} \le \frac{h}{W(I)}\int_0^\infty (I + Q^*)|\rho_t(s) - \rho(s)|ds + \frac{b}{W(I)}\int_0^\infty s|\rho_t(s) - \rho(s)|ds$$

$$\le \frac{MG(I)}{W(I)}\int_0^\infty |\rho_t(s) - \rho(s)|ds + b2^{\frac{1}{2}}M'\left(\int_0^\infty |\rho_t(s) - \rho(s)|ds\right)^{1/2}.$$

Finally, taking expectation, (28) and Property D.4 prove the result.

(c) For each $I \in \mathbb{I}$ and $t \geq 0$, by adding and subtracting the term

$\alpha \int_0^\infty V_{t-1}\big((I+q-s)^+\big)\rho(s)ds$, we have

$$|V_t(I) - V^*(I)| \leq \sup_{q \in \mathbb{Q}^*} |c_t(I,q) - (I,q)| + \sup_{q \in \mathbb{Q}^*} \alpha \int_0^\infty V_{t-1}\big((I+q-s)^+\big)|\rho_t(s) - \rho(s)|ds$$

$$+ \alpha \int_0^\infty \big|V_{t-1}\big((I+q-s)^+\big) - V^*\big((I+q-s)^+\big)\big|\rho(s)ds$$

$$\leq \sup_{q \in \mathbb{Q}^*} |c_t(I,q) - (I,q)| + \frac{\alpha}{1-\alpha\beta} \sup_{q \in \mathbb{Q}^*} \int_0^\infty W\big((I+q-s)^+\big)|\rho_t(s) - \rho(s)|ds$$

$$+ \alpha\beta\|V_{t-1} - V^*\|_W W(I),$$

where the last inequality is due to (35), (17), (14), and (15). Therefore, from (15) and (19) and by taking expectation,

$$E\|V_t - V^*\|_W \leq E \sup_{q \in \mathbb{Q}^*} |c_t(I,q) - c(I,q)| + \frac{\alpha}{1-\alpha\beta} E\|\rho_t - \rho\| + \alpha\beta E\|V_{t-1} - V^*\|_W. \tag{50}$$

Finally, from (17) and (35), $\eta := \limsup_{t \to \infty} E\|V^* - V_t\|_W < \infty$. Hence, taking limsup in both sides of (50), from part (a) and property D.5 in Theorem 4.1, we get $\eta \leq \alpha\beta\eta$, which yields $\eta = 0$ (since $0 < \alpha\beta < 1$). This proves (c).

(d) For each $t \geq 0$, let $H_t : \mathbb{I} \to \mathfrak{R}$ be the function defined as

$$H_t(y) := cy + L_t(y) + \alpha \int_0^\infty V_{t-1}\big((y-s)^+\big)\rho_t(s)ds.$$

Hence, (34) is equivalent to

$$V_t(I) = \min_{q \in \mathbb{Q}^*} H_t(I+q) - cI, \qquad I \in \mathbb{I}. \tag{51}$$

Moreover (see (33)), observe that H_t is convex and $\lim_{y \to \infty} H_t(y) = \infty$. Thus, there exist a constant $K_t \geq 0$ such that

$$H_t(K_t) = \min_{I \leq y \leq Q^* + I} H_t(y),$$

and

$$g_t(I) = \begin{cases} K_t - I & \text{if} \quad 0 \leq I \leq K_t \\ 0 & \text{if} \quad I > K_t \end{cases}$$

attains the minimum in (51). $\blacksquare$

6.3 Proof of Theorem 5.3

We fix an arbitrary $I \in \mathbb{I}$. Since $g_\infty(I)$ is an accumulation point of $\{g_t(I)\}$ (see Remark 5.2), there exists a subsequence $\{t_m(I)\}$ of $\{t\}$ $(t_m = t_m(I))$ such that

$$g_{t_m}(I) \to g_\infty(I) \quad \text{as } m \to \infty.$$

Moreover, from (34) and Theorem 5.1(d), letting $t_m = m,$ we have

$$V_m(I) = c_m(I, g_m) + \alpha \int_0^\infty V_{m-1}\left(\left(I + g_m - s\right)^+\right)\rho_m(ds). \qquad (52)$$

On the other hand, following similar arguments as the proof of Theorem 5.1(c), for each $m \geq 0$ and $(I, q) \in \mathbb{I} \times \mathbb{Q},$ we have

$$\left| \alpha \int_0^\infty V_{m-1}((I + q - s)^+)\rho_m(s)ds - \alpha \int_0^\infty V^*\left((I + q - s)^+\right)\rho(s)ds \right|$$

$$\leq \alpha \int_0^\infty \left| V_{m-1}((I + q - s)^+) - V^*\left((I + q - s)^+\right) \right| \rho_m(s)ds + \alpha \int_0^\infty V^*\left((I + q - s)^+\right)|\rho_m(s) - \rho(s)|ds$$

$$\leq \alpha\beta \| V_{m-1} - V^* \|_W W(I) + \frac{\alpha}{1 - \alpha\beta} \|\rho_m - \rho\|.$$

Then, for each $I \in \mathbb{I},$

$$E \sup_{q \in \mathbb{Q}^*} \left| \alpha \int_0^\infty V_{m-1}((I + q - s)^+)\rho_m(s)ds - \alpha \int_0^\infty V^*\left((I + q - s)^+\right)\rho(s)ds \right| \to 0, \quad \text{as } m \to \infty.$$

$$(53)$$

Now,

$$\alpha \int_0^\infty V_{m-1}\left(\left(I + g_m - s\right)^+\right)\rho_m(ds)$$

$$= \left[\alpha \int_0^\infty V_{m-1}\left(\left(I + g_m - s\right)^+\right)\rho_m(ds) - \alpha \int_0^\infty V^*\left(\left(I + g_m - s\right)^+\right)\rho(s)ds \right] \qquad (54)$$

$$+ \alpha \int_0^\infty V^*\left(\left(I + g_m - s\right)^+\right)\rho(s)ds.$$

Taking expectation and liminf as $m \to \infty$ on both sides of (54), from (53) we obtain

$$\liminf_{m \to \infty} E\alpha \int_0^\infty V_{m-1}\left(\left(I + g_m - s\right)^+\right)\rho_m(ds) = \liminf_{m \to \infty} E\alpha \int_0^\infty V^*\left(\left(I + g_m - s\right)^+\right)\rho(s)ds$$

$$\geq \int_0^\infty V^*\left(\left(I + g_\infty - s\right)^+\right)\rho(s)ds,$$

where the last inequality follows by applying Fatou's Lemma and because the function $q \to (I + q - s)^+$ is continuous. Hence, taking expectation and liminf in (52), we obtain

$$c\left(I, g_{\infty}\right) + \alpha \int_0^{\infty} V^* \left(\left(I + g_{\infty} - s\right)^+\right) \rho(s) ds \leq V^*(I), \quad I \in \mathbb{I}. \qquad (55)$$

As I was arbitrary, by (18), the equality holds in (55) for all $I \in \mathbb{I}$. To conclude, standard arguments on stochastic control literature (see, e.g., [10]) show that the policy $\gamma^* = \{g_{\infty}\}$ is optimal. ∎

7. Concluding remarks

In this chapter we have introduced an estimation and control procedure in inventory systems when the density of the demand is unknown by the inventory manager. Specifically we have proposed a density estimation method defined by the projection to a suitable set of densities, which, combined with control schemes relative to the inventory systems, defines a procedure to construct optimal ordering policies.

A point to highlight is that our results include the most general scenarios of an inventory system, e.g., state and control spaces either countable or uncountable, possibly unbounded costs, finite or infinite inventory capacity. This generality entailed the need to develop new estimation and control techniques, accompanied by a suitable mathematical analysis. For example, the simple fact of considering possibly unbounded costs led us to formulate a density estimation method that was related to the weight function W, which, in turn, defines the normed linear space $\mathbb{B}_W$ (see (15)), all this through the projection estimator. Observe that if the cost function c is bounded, we can take $W \equiv 1$ and we have $\|\cdot\| = \|\cdot\|_{L_1}$ (see (19) and (25)). Thus, any L_1−consistent density estimator ρ_t can be used for the construction of optimal ordering policies.

Finally, the theory presented in this chapter lays the foundations to develop estimation and control algorithms in inventory systems considering other optimality criteria, for instance, the average cost or discounted criteria with random state-action-dependent discount factors (see [14, 15] and references therein).

Author details

Jesús Adolfo Minjárez-Sosa
Departamento de Matemáticas, Universidad de Sonora, Hermosillo, Sonora, Mexico

*Address all correspondence to: aminjare@gauss.mat.uson.mx

IntechOpen

References

[1] Arrow KJ, Karlin S, Scarf H. Studies in the Mathematical Theory of Inventory and Production. CA: Stanford University Press; 1958

[2] Bensoussan A, Çakanyıldırım M, Sethi SP. Partially observed inventory systems: The case of zero balance walk. SIAM Journal on Control and Optimization. 2007;**46**:176-209

[3] Bensoussan A, Çakanyıldırım M, Minjárez-Sosa JA, Royal A, Sethi SP. Inventory problems with partially observed demands and lost sales. Journal of Optimization Theory and Applications. 2008;**136**:321-340

[4] Bensoussan A, Çakanyıldırım M, Minjárez-Sosa JA, Sethi SP, Shi R. Partially observed inventory systems: The case of rain checks. SIAM Journal on Control and Optimization. 2008; **47**(5):2490-2519

[5] Bensoussan A, Çakanyıldırım M, Minjárez-Sosa JA, Sethi SP, Shi R. An incomplete information inventory model with presence of inventories or backorders as only observations. Journal of Optimization Theory and Applications. 2010;**146**(3):544-580

[6] Bertsekas DP. Dynamic Programming: Deterministic and Stochastic Models. Englewood Cliffs, N. J: Prentice-Hall; 1987

[7] Beyer D, Cheng F, Sethi SP, Taksar MI. Markovian Demand Inventory Models. New York: Springer; 2008

[8] Dynkin EB, Yushkevich AA. Controlled Markov Processes. New York: Springer-Verlag; 1979

[9] Gordienko EI, Minjárez-Sosa JA. Adaptive control for discrete-time Markov processes with unbounded costs: Discounted criterion. Kybernetika. 1998;**34**:217-234

[10] Hernández-Lerma O, Lasserre JB. Discrete-Time Markov Control Processes: Basic Optimality Criteria. New York: Springer-Verlag; 1996

[11] Hernández-Lerma O, Lasserre JB. Further Topics on Discrete-Time Markov Control Processes. New York: Springer-Verlag; 1999

[12] Hilgert N, Minjárez-Sosa JA. Adaptive policies for time-varying stochastic systems under discounted criterion. Mathematical Methods of Operations Research. 2001;**54**(3): 491-505

[13] Minjárez-Sosa JA. Approximation and estimation in Markov control processes under discounted criterion. Kybernetika. 2004;**6**(40):681-690

[14] Minjárez-Sosa JA. Empirical estimation in average Markov control processes. Applied Mathematics Letters. 2008;**21**:459-464

[15] Minjárez-Sosa JA. Markov control models with unknown random state-action-dependent discount factors. TOP. 2015;**23**:743-772

A Comparative Study of Maximum Likelihood Estimation and Bayesian Estimation for Erlang Distribution and Its Applications

Kaisar Ahmad and Sheikh Parvaiz Ahmad

Abstract

In this chapter, Erlang distribution is considered. For parameter estimation, maximum likelihood method of estimation, method of moments and Bayesian method of estimation are applied. In Bayesian methodology, different prior distributions are employed under various loss functions to estimate the rate parameter of Erlang distribution. At the end the simulation study is conducted in R-Software to compare these methods by using mean square error with varying sample sizes. Also the real life applications are examined in order to compare the behavior of the data sets in the parametric estimation. The comparison is also done among the different loss functions.

Keywords: Erlang distribution, prior distributions, loss functions, simulation study, applications

1. Introduction

Erlang distribution is a continuous probability distribution with wide applicability, primarily due to its relation to the exponential and gamma distributions. The Erlang distribution was developed by Erlang [1] to examine the number of telephone calls that could be made at the same time to switching station operators. This distribution can be expressed as waiting time and message length in telephone traffic. If the duration of individual calls are exponentially distributed then the duration of succession of calls is the Erlang distribution. The Erlang variate becomes gamma variate when its shape parameter is an integer (for details see Evans et al. [2]). Bhattacharyya and Singh [3] obtained Bayes estimator for the Erlangian queue under two prior densities. Haq and Dey [4] addressed the problem of Bayesian estimation of parameters for the Erlang distribution assuming different independent informative priors. Suri et al. [5] used Erlang distribution to design a simulator for time estimation of project management process. Damodaran et al. [6] obtained the expected time between failure measures. Further, they showed that the predicted failure times are closer to the actual failure times. Jodra [7] showed the

procedure of computing the asymptotic expansion of the median of Erlang distribution.

The probability density function of an Erlang variate is given by

$$f(x; \lambda, k) = \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} \quad for \, x > 0, \, k \in N \, and \, \lambda > 0. \tag{1}$$

where λ and k are the rate and the shape parameters, respectively, such that is k an integer number.

1.1 Graphical representation of pdf for Erlang distribution

In this chapter, Erlang distribution is considered. Some structural properties of Erlang distribution have been obtained. The parameter estimation of Erlang distribution is obtained by employing the maximum likelihood method of estimation, method of moments and Bayesian method of estimation in different sections of this chapter. In Bayesian approach, the parameters are estimated by using Jeffrey's and Quasi priors under different loss functions (**Figure 1**).

1.2 Relationship of Erlang distribution with other distributions

   i. The gamma distribution is a generalized form of the Erlang distribution.

   ii. If the shape parameter k is 1, then Erlang distribution reduces to exponential distribution.

   iii. If the scale parameter is 2, then Erlang distribution reduces to Chi-square distribution with 2 degrees of freedom.

Thus from the above descriptions, we can say that exponential distribution and Chi-square distribution are the sub-models of Erlang distribution.
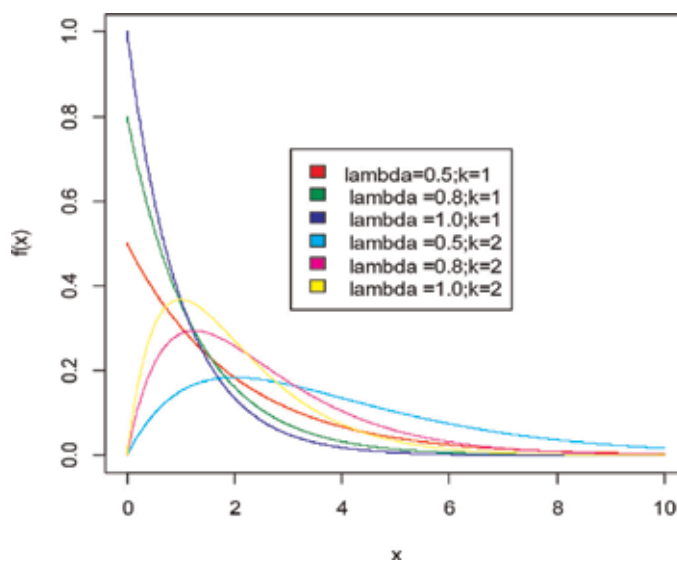
Figure 1.
Pdf's of Erlang distribution for different values of lambda and k.

2. Methods used for parameter estimation

In this chapter, we have used different approaches for parameter estimation. The first two methods come under the classical approach which was founded by Fisher in a series of fundamental papers round about 1930.

The alternative approach is the Bayesian approach which was first discovered by Reverend Thomas Bayes. In this chapter, we have used two different priors for the parameter estimation. Also three loss functions are used which are discussed in their respective sections. A number of symmetric and asymmetric loss functions used by various researchers; see Zellner [8], Ahmad and Ahmad [9], Ahmad et al. [10], etc.

These methods of estimations are elaborated in their respective sections accordingly.

2.1 Maximum likelihood (MLH) estimation

The most general method of estimation is known as maximum likelihood (MLH) estimators, which was initially formulated by Gauss. Fisher in the early 1920 firstly introduced MLH as general method of estimation and later on developed by him in a series of papers. He revealed the advantages of this method by showing that it yields sufficient estimators, which are asymptotically MVUES. Thus the important feature of this method is that we look at the value of the random sample and then select our estimate of the unknown population parameter, the value of which the probability of getting the observed data is maximum.

Suppose the observed data sample values are $(x_1, x_2, ..., x_n)$. When X is a discrete random variable, we can write $P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = f(x_1, x_2, ..., x_n)$, which is the value of joint probability distribution at the sample point $(x_1, x_2, ..., x_n)$. Since the sample values has been observed and are therefore fixed numbers, we consider $f(x_1, x_2, ..., x_n; \lambda)$ as the value of a function of the parameter λ, referred to as the likelihood function.

Similarly the definition applies when the random sample comes from a continuous population but in that case $f(x_1, x_2, ..., x_n; \lambda)$ is the value of joint pdf at the sample point $(x_1, x_2, ..., x_n)$. That is, the likelihood function at the sample value $(x_1, x_2, ..., x_n)$ which is given by

$$L(x_1, x_2, ..., x_n | \lambda) = \prod_{i=1}^{m} f(x_i, \lambda).$$

Since the principle of maximum likelihood consists in finding an estimator of the parameter which maximizes the likelihood function for variation in the parameter. Thus if there exists a function $\hat{\lambda} = \hat{\lambda}(x_1, x_2, ..., x_n)$ of the sample values which maximizes $L(x|\lambda)$ for variation in λ, then $\hat{\lambda}$ is to be taken as the estimator of λ. Usually we call $\hat{\lambda}$ as ML estimators. Thus $\hat{\lambda}$ is the solution

$$iff \quad \frac{\partial L(x|\lambda)}{\partial \lambda} = 0 \ and \ \frac{\partial^2 L(x|\lambda)}{\partial \lambda^2} < 0.$$

Since $L(x|\lambda) > 0$, so $\log L(x|\lambda)$ which shows that $L(x|\lambda)$ and $\log L(x|\lambda)$ attains their extreme values at $\hat{\lambda}$. Therefore, the equation becomes

$$\frac{1}{L(x|\lambda)} \frac{\partial L(x|\lambda)}{\partial \lambda} = 0 \ \Rightarrow \ \frac{\partial \log L(x|\lambda)}{\partial \lambda} = 0$$

a form which is more convenient from practical point of view.

The MLH estimation of the rate parameter of Erlang distribution is obtained in the following theorem:

Theorem 2.1: Let $(x_1, x_2, ..., x_n)$ be a random sample of size n from Erlang density function Eq. (1), then the maximum likelihood estimator of λ is given by

$$\hat{\lambda} = \frac{nk}{\sum_{i=1}^{n} x_i}.$$

Proof: The likelihood function of random sample of size n having Erlang density function Eq. (1) is given by

$$L(x; \lambda, k) = \left(\frac{(\lambda)^k}{(k-1)!} \right)^n \prod_{i=1}^{n} x_i^{k-1} e^{-\lambda \sum_{i=1}^{n} x_i}. \tag{2}$$

The log likelihood function is given by

$$\log L(x; \lambda, \beta) = nk \log \lambda + (k-1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i - n \log (k-1)!. \tag{3}$$

Differentiating Eq. (3) w.r.t. λ and equating to zero, we get

$$\frac{\partial}{\partial \lambda} \left(nk \log \lambda + (k-1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i - n \log (k-1)!. \right) = 0$$

$$\hat{\lambda} = \frac{nk}{\sum_{i=1}^{n} x_i}. \tag{4}$$

2.2 Method of moments (MM)

One of the simplest and oldest methods of estimation is the method of moments. The method of moments was discovered by Karl Pearson in 1894. It is a method of estimation of population parameters such as mean, variance, etc. (which need not be moments), by equating sample moments with unobservable population moments and then solving those equations for the quantities to be estimated. The method of moments is special case when we need to estimate some known function of finite number of unknown moments.

Suppose $f(x; \lambda_1, \lambda_2, ..., \lambda_p)$ be the density function of the parent population with p parameters $(\lambda_1, \lambda_2, ..., \lambda_p)$. Let μ'_s be the sth moment of a random variable about origin and is given by

$$\mu'_s = \int_{-\infty}^{\infty} x^s f(x; \lambda_1, \lambda_2, ..., \lambda_p) \ ; r = 1, 2, ..., p.$$

In general $\left(\mu'_1, \mu'_2, ..., \mu'_p \right)$ will be the functions of parameters $(\lambda_1, \lambda_2, ..., \lambda_p)$. Let $(x_i; i = 1, 2, ..., n)$ be a random sample of size n from the given population. The method of moments consists in solving the p-equations (i) for $(\lambda_1, \lambda_2, ..., \lambda_p)$ in terms of $\left(\mu'_1, \mu'_2, ..., \mu'_p \right)$. Then replacing these moments $(\mu'_s; s = 1, 2, 3, ..., p)$ by the sample moments

e.g., $\hat{\lambda}_i = \hat{\lambda} \left(\mu'_1, \mu'_2, ..., \mu'_p \right) = \lambda_i \left(m'_1, m'_2, ..., m'_p \right) \ ; i = 1, 2, ..., p.$

where m_i is the ith moment about origin in the sample.

Then by the method of moments $\left(\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p\right)$ are the estimators of respectively.

The MM estimation of the rate parameter of Erlang distribution is obtained in the following theorem:

Theorem 2.2: Let $(x_1, x_2, ..., x_n)$ be a random sample of size n from Erlang density function Eq. (1), then the moment estimator of λ is given by

$$\hat{\lambda} = \left(\frac{k}{\bar{x}}\right).$$

Proof: If the numbers $(x_1, x_2, ..., x_n)$ represents a set of data, then an unbiased estimator for the rth moment about origin is

$$\hat{m}_r = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

where $\hat{m}_r$ stands for the estimate of m_r.

The rth moment of two parameter Erlang distribution about origin is given by

$$\mu'_r = \int_0^\infty x^r f(x; \lambda, k) dx \tag{6}$$

Using Eq. (5) in Eq. (6), we have

$$\mu'_r = \frac{\Gamma(r+k)}{\lambda^r (k-1)!}. \tag{7}$$

If r = 1 in Eq. (7), we get

$$\mu'_1 = \frac{k}{\lambda}.$$

If r = 2, then Eq. (7) becomes

$$\mu'_2 = \frac{k(k+1)}{\lambda^2}.$$

Thus the variance is given by $\sigma^2 = \frac{k}{\lambda^2}$.

When we divide σ^2 by μ'^2_1, we get an expression which is a function of k only and is given by

$$\frac{\sigma^2}{\mu'^2_1} = \frac{\left(\frac{k}{\lambda^2}\right)}{\left(\frac{k}{\lambda}\right)^2}$$

$$\frac{\sigma^2}{\mu'^2_1} = \frac{1}{k}. \tag{8}$$

On taking the square roots of Eq. (8), we have the coefficient of variation

$$\frac{\sigma}{\mu'_1} = \sqrt{\left(\frac{1}{k}\right)}. \tag{9}$$

The rate parameter λ can be then estimated using the following equation

$$m_1' = \mu_1'.$$

If r = 1, in Eq. (5), then

$$m_1' = \overline{x}.$$

Also if r = 1 in Eq. (7), then

$$\mu_1' = \frac{k}{\lambda}.$$

Thus,

$$m_1' = \mu_1'$$

$\overline{x} = \frac{k}{\lambda}$, where $\overline{x}$ is the mean of the data and

$$\hat{\lambda} = \left(\frac{k}{\overline{x}}\right).$$

3. Bayesian method of estimation

Nowadays, the Bayesian school of thought is garnering more attention and at an increasing rate. This thought of statistics was given by Reverend Thomas Bayes. He first discovered the theorem that now bears his name. It was written up in a paper "An Essay Towards Solving a Problem in the Doctrine of Chances." This paper was found after his death by his friend Richard Price, who had it published posthumously in the Philosophical Transactions of the Royal Society in 1763. Bayes showed how inverse probability could be used to calculate probability of antecedent events from the occurrence of the consequent event. His methods were adopted by Laplace and other scientists in the nineteenth century. By mid twentieth century interest in Bayesian methods was renewed by De Finetti, Jeffreys and Lindley, among others. They developed a complete method of statistical inference based on Bayes' theorem. Bayesian analysis is to be used by practitioners for situations where scientists have a priori information about the values of the parameters to be estimated. In everyday life, uncertainty often permeates our choices, and when choices need to be made, past experience frequently proves a helpful aid. Bayesian theory provides a general and consistent framework for dealing with uncertainty.

Within Bayesian inference, there are also different interpretations of probability, and different approaches based on those interpretations. Early efforts to make Bayesian methods accessible for data analysis were made by Raiffa and Schlaifer [11], DeGroot [12], Zellner [13], and Box and Tiao [14]. The most popular interpretations and approaches are objective Bayesian inference and subjective Bayesian inference. Excellent expositions of these approaches are with Bayes and Price [15], Laplace [16], Jeffrey's [17], Anscombe and Aumann [18], Berger [19, 20], Gelman et al. [21], Leonard and Hsu [22], De-Finetti [23]. Modern Bayesian data analysis and methods based on Markov chain Monte Carlo methods are presented in Bernardo and Smith [24], Robert [25], Gelman et al. [26], Marin and Robert [27], Carlin and Louis [28]. Good elementary introductions to the subject are Ibrahim et al. [29], Ghosh [30], Bansal [31], Koch [32], Hoff [33].

In Bayesian statistics probability is not defined as a frequency of occurrence but as the plausibility that a proposition is true, given the available information. The parameters are treated as random variables. The rules of probability are used

directly to make inferences about the parameters. Probability statements about parameters must be interpreted as "degree of belief." We revise our beliefs about parameters after getting the data by using Bayes' theorem. This gives our posterior distribution which gives the relative weights we give to each parameter value after analyzing the data. The posterior distribution comes from two sources: the prior distribution and the observed data. This means that the inference is based on the actual occurring data, not all possible data sets that might have occurred.

In this section, posterior distribution of Erlang distribution is obtained by using Jeffrey's prior and Quasi prior. The rate parameter of Erlang distribution is estimated with the help of different loss functions. For parameter estimation we have used the approach as is used by Ahmad et al. [34], Ahmad et al. [35], etc. Some important prior distributions and loss functions which we have used in this article are given as below:

3.1 Prior distributions used

Prior distribution is the basic part of Bayesian analysis which represents all that is known or assumed about the parameter. Usually the prior information is subjective and is based on a person's own experience and judgment, a statement of one's degree of belief regarding the parameter.

Another important feature of the Bayesian analysis is the choice of the prior distribution. If the data have sufficient signal, even a bad prior will still not greatly influence the posterior. We can examine the impact of prior by observing the stability of posterior distribution related to different choices of priors. If the posterior distribution is highly dependent on the prior, then the data (the likelihood function) may not contain sufficient information. However, if the posterior is relatively stable over a choice of priors, then the data indeed contains significant information.

Prior distribution may be categorical in different ways. One common classification is a dichotomy that separated "proper" and "improper" priors.

A prior distribution is proper if it does not depend on the data and the value of integral $\int_{-\infty}^{\infty} g(\lambda)d\lambda$ or summation $\sum g(\lambda)$ is one. If the prior does not depend on the data and the distribution does not integrate or sum to one then we say that the prior is improper.

In this chapter, we have used two different priors Jeffrey's prior and Quasi prior which are given below:

3.1.1 Jeffrey's prior

An invariant form for the prior probability in estimation problems is given by Jeffery's [36].

The general formula of the Jeffreys prior, which is defined by

$$g(\lambda) \propto \sqrt{I(\lambda)} \propto \left(-E\left[\frac{\partial^2 \log L(\lambda|x)}{\partial \lambda^2} \right] \right)^{\frac{1}{2}}.$$

where $I(\lambda)$ is the Fisher information for the parameter λ. When there are multiple parameters I is the Fisher information matrix, the matrix of the expected second partials

$$I(\lambda) = E\left(\frac{\partial^2 \log L(\lambda|x)}{\partial \lambda_i \partial \lambda_j}\right).$$

In this situation, the Jeffreys prior is given by

$$g(\lambda) \propto \sqrt{\det(I(\lambda))}.$$

Jeffrey suggested a thumb rule for specifying non-informative prior for parameter λ as

Rule 1: if $\lambda \in (-\infty, \infty)$ take $g(\lambda)$ to be constant, i.e., λ to be uniformly distributed.

Rule 2: if $\lambda \in (0, \infty)$ take $g(\lambda) \propto \frac{1}{\lambda}$, i.e., $\log \lambda$ to be uniformly distributed.

Under linear transformation, rule 1 is invariant and under any power transformation of λ, rule 2 is invariant.

3.1.2 Quasi prior

When there is no more information about the distribution parameter, one may use the Quasi density as given by $g(\lambda) = \frac{1}{\lambda^d}$; $\lambda > 0$ *and* $d > 0$.

3.2 Loss functions used

The concept of loss function is old as Laplace and was reintroduced in statistics by Abraham Wald [37]. In statistics, typically a loss function is used for parameter estimation, and the event in question is some function of the difference between estimated and true values for an occurrence of data. In the context of economics, loss function is usually economic cost. In optimal control, the loss is the penalty for failing to achieve a desired value.

The word "loss" is used in place of "error" and the loss function is used as a measure of the error or loss. Loss function is a measure of the error and presumably would be greater for large error than for small error. We would want the loss to be small or we want the estimate to be close to what it is estimating. Loss depends on sample and we cannot hope to make the loss small for every possible sample but can try to make the loss small on the average. Our objective is to select an estimator that makes this error or loss small, also which makes the average loss (risk) small and ideally select an estimator that has the small risk.

In this chapter, we have used three different Loss Functions which are as under:

3.2.1 Precautionary loss function (PLF)

The concept of precautionary loss function (PLF) was introduced by Norstrom [38]. He introduced an alternative asymmetric loss function and also presented a general class of precautionary loss functions as a special case. These loss functions approach infinitely near the origin to prevent underestimation, thus giving conservative estimators, especially when low failure rates are being estimated. These estimators are very useful when underestimation may lead to serious consequences. A very useful and simple asymmetric precautionary loss function (PLF) is

$$L\left(\hat{\lambda}, \lambda\right) = \frac{\left(\hat{\lambda} - \lambda\right)^2}{\lambda}.$$

3.2.2 Al-Bayyati's loss function (ALF)

The loss function proposed by Al-Bayyati [39] is an asymmetric loss function and is given by

$$l_A\left(\hat{\lambda}, \lambda\right) = \lambda^c \left(\hat{\lambda} - \lambda\right)^2 c \varepsilon R.$$

where λ and $\hat{\lambda}$ represents the true and estimated values of the parameter. This loss function is frequently used because of its analytical tractability in Bayesian analysis.

3.2.3 LINEX loss function (LLF)

The idea of LINEX loss function (LLF) was founded by Klebanov [40] and used by Varian [41] in the context of real estate assessment. The formula of LLF is given by

$$L\left(\hat{\lambda}, \lambda\right) = \left(\exp\left(a\left(\hat{\lambda} - \lambda\right)\right) - a\left(\hat{\lambda} - \lambda\right) - 1\right)$$

where λ and $\hat{\lambda}$ represents the true and estimated values of the parameter and the constant c determines the shape of the loss function.

3.3 Posterior density under Jeffrey's prior

Let $(x_1, x_2, ..., x_n)$ be a random sample of size n having the Erlang density function Eq. (1) which is given by

$$f(x; \lambda, k) = \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} \quad for\, x > 0, k \in N\, and\, \lambda > 0$$

and the likelihood function Eq. (2) given as below

$$L(x; \lambda, k) = \left(\frac{(\lambda)^k}{(k-1)!}\right)^n \prod_{i=1}^{n} x_i^{k-1} e^{-\lambda \sum_{i=1}^{n} x_i}.$$

Jeffreys' non-informative prior for λ is given by

$$g(\lambda) \propto \sqrt{\det(I(\lambda))}.$$

where $I(\lambda) = -nE\left[\frac{\partial^2 \log f(x; \lambda, k)}{\partial \lambda^2}\right]$ is the Fisher's information matrix for the probability density function Eq. (1).

On solving the above expression, we have

$$g(\lambda) = \frac{1}{\lambda}. \tag{10}$$

By using the Bayes theorem, we have

$$\pi_1\left(\lambda | \underline{x}\right) \propto L(x|\lambda) g(\lambda). \tag{11}$$

Using Eqs. (2) and (10) in Eq. (11), we get

$$\pi_1\left(\lambda|\underline{x}\right) \propto \frac{(\lambda)^{nk-1}}{(k-1)!}\prod_{i=1}^{n}x_i^{\,k-1}e^{-\lambda\sum_{i=1}^{n}x_i}$$

$$\pi_1\left(\lambda|\underline{x}\right) = \rho\lambda^{nk-1}e^{-\lambda\sum_{i=1}^{n}x_i} \tag{12}$$

where ρ is independent of λ and

$$\rho^{-1} = \int_{0}^{\infty}\lambda^{nk-1}e^{-\lambda\sum_{i=1}^{n}x_i}\,d\lambda$$

$$\rho = \frac{\left(\sum_{i=1}^{n}x_i\right)^{nk}}{\Gamma nk}.$$

Using the value of ρ in Eq. (12)

$$\pi_1\left(\lambda|\underline{x}\right) = \left(\frac{\lambda^{nk-1}e^{-\lambda\sum_{i=1}^{n}x_i}\left(\sum_{i=1}^{n}x_i\right)^{nk}}{\Gamma nk}\right). \tag{13}$$

3.4 Posterior density under Quasi prior

Let $(x_1, x_2, ..., x_n)$ be a random sample of size n having the Erlang density function Eq. (2) and the likelihood function Eq. (2).

The Quasi for λ is given by

$$g(\lambda) = \frac{1}{\lambda^d}. \tag{14}$$

By using the Bayes theorem, we have

$$\pi_2\left(\lambda|\underline{x}\right) \propto L(x|\lambda)g(\lambda). \tag{15}$$

Using Eqs. (2) and (14) in Eq. (15), we have

$$\pi_2\left(\lambda|\underline{x}\right) \propto \frac{(\lambda)^{nk-d}}{(k-1)!}\prod_{i=1}^{n}x_i^{\,k-1}e^{-\lambda\sum_{i=1}^{n}x_i}$$

$$\pi_2\left(\lambda|\underline{x}\right) = \rho\lambda^{nk-d}e^{-\lambda\sum_{i=1}^{n}x_i} \tag{16}$$

where ρ is independent of λ and

$$\rho^{-1} = \int_{0}^{\infty}\lambda^{nk-2c}e^{-\lambda\sum_{i=1}^{n}x_i}\,d\lambda$$

$$\rho = \frac{\left(\sum_{i=1}^{n}x_i\right)^{nk-d+1}}{\Gamma(nk-d+1)}.$$

By using the value of ρ in Eq. (16), we have

$$\pi_2\left(\lambda|\underline{x}\right) = \left(\frac{\lambda^{nk-d}e^{-\lambda\sum\limits_{i=1}^{n}x_i}\left(\sum_{i=1}^{n}x_i\right)^{nk-d+1}}{\Gamma(nk-d+1)}\right). \tag{17}$$

4. Estimation of parameters under Jeffrey's prior

In this section, parameter estimation of Erlang distribution is done by using Jeffreys' prior under different loss functions. The procedure of calculating the Bayesian estimate is already defined in Section 3. The estimates are obtained in the following theorems:

Theorem 4.1: Assuming the loss function $L_p\left(\hat{\lambda},\lambda\right)$, the Bayesian estimator of the rate parameter λ, if the shape parameter k is known, is of the form

$$\hat{\lambda}_p = \frac{\sqrt{nk(nk+1)}}{\left(\sum_{i=1}^{n}x_i\right)}.$$

Proof: The risk function of the estimator λ under the precautionary loss function $L_p\left(\hat{\lambda},\lambda\right)$ is given by the formula

$$R\left(\hat{\lambda}\right) = \int_0^\infty \frac{\left(\hat{\lambda}-\lambda\right)^2}{\hat{\lambda}}\pi_1\left(\lambda|\underline{x}\right)d\lambda. \tag{18}$$

Using Eq. (13) in Eq. (18), we get

$$R\left(\hat{\lambda}\right) = \int_0^\infty \frac{\left(\hat{\lambda}-\lambda\right)^2}{\hat{\lambda}}\left(\frac{\lambda^{nk-1}\left(\sum_{i=1}^{n}x_i\right)^{nk}e^{-\lambda\sum\limits_{i=1}^{n}x_i}}{\Gamma nk}\right)d\lambda$$

$$R\left(\hat{\lambda}\right) = \frac{\left(\sum_{i=1}^{n}x_i\right)^{nk}}{\Gamma nk}\left[\hat{\lambda}\int_0^\infty \lambda^{nk-1}e^{-\lambda\sum\limits_{i=1}^{n}x_i}d\lambda + \frac{1}{\hat{\lambda}}\int_0^\infty \lambda^{nk+1}e^{-\lambda\sum\limits_{i=1}^{n}x_i}d\lambda - 2\int_0^\infty \lambda^{nk}e^{-\lambda\sum\limits_{i=1}^{n}x_i}d\lambda\right].$$

On solving the above expression, we get

$$R\left(\hat{\lambda}\right) = \hat{\lambda} + \frac{1}{\hat{\lambda}}\frac{nk(nk+1)}{\left(\sum_{i=1}^{n}x_i\right)^2} - \frac{2nk}{\left(\sum_{i=1}^{n}x_i\right)}.$$

Minimization of the risk with respect to $\hat{\lambda}$ gives us the optimal estimator i.e.,

$$\frac{\partial}{\partial\hat{\lambda}}\left[R\left(\hat{\lambda}\right)\right] = 0$$

$$\hat{\lambda}_p = \frac{\sqrt{nk(nk+1)}}{\left(\sum_{i=1}^{n}x_i\right)}. \tag{19}$$

Theorem 4.2: Assuming the loss function $l_A\left(\hat{\lambda},\lambda\right)$, the Bayesian estimator of the rate parameter λ, if the shape parameter k is known, is of the form

$$\hat{\lambda}_A = \frac{(nk+c)}{\left(\sum_{i=1}^{n}x_i\right)}.$$

Proof: The risk function of the estimator λ under the Al-Bayyati's loss function $L_A\left(\hat{\lambda}, \lambda\right)$ is given by the formula

$$R\left(\hat{\lambda}\right) = \int_0^\infty \lambda^c \left(\hat{\lambda} - \lambda\right)^2 \pi_1\left(\lambda | \underline{x}\right) d\lambda. \tag{20}$$

On substituting Eq. (13) in Eq. (20), we have.

$$R\left(\hat{\lambda}\right) = \int_0^\infty \lambda^c \left(\hat{\lambda} - \lambda\right)^2 \left(\frac{\lambda^{nk-1}\left(\sum_{i=1}^n x_i\right)^{nk} e^{-\lambda \sum_{i=1}^n x_i}}{\Gamma nk} \right) d\lambda$$

$$R\left(\hat{\lambda}\right) = \frac{\left(\sum_{i=1}^n x_i\right)^{nk}}{\Gamma nk} \left[\hat{\lambda}^2 \int_0^\infty \lambda^{nk+c-1} e^{-\lambda \sum_{i=1}^n x_i} d\lambda + \int_0^\infty \lambda^{nk+c+1} e^{-\lambda \sum_{i=1}^n x_i} d\lambda - 2\hat{\lambda} \int_0^\infty \lambda^{nk+c} e^{-\lambda \sum_{i=1}^n x_i} d\lambda \right].$$

On solving the above expression, we get

$$R\left(\hat{\lambda}\right) = \frac{\hat{\lambda}^2 \Gamma(nk+c)}{\Gamma nk \left(\sum_{i=1}^n x_i\right)^c} + \frac{\Gamma(nk+c+2)}{\Gamma nk \left(\sum_{i=1}^n x_i\right)^{c+2}} - \frac{2\hat{\lambda}\Gamma(nk+c+1)}{\Gamma nk \left(\sum_{i=1}^n x_i\right)^{c+1}}.$$

Minimization of the risk with respect to $\hat{\lambda}$ gives us the optimal estimator i.e.,

$$\frac{\partial}{\partial \hat{\lambda}} \left[R\left(\hat{\lambda}\right)\right] = 0$$

$$\hat{\lambda}_A = \frac{(nk+c)}{\left(\sum_{i=1}^n x_i\right)}. \tag{21}$$

Theorem 4.3: Assuming the loss function $L_l\left(\hat{\lambda}, \lambda\right)$, the Bayesian estimator of the rate parameter λ, if the shape parameter k is known, is of the form

$$\hat{\lambda}_l = \frac{nk \log\left(1 + \frac{a}{\left(\sum_{i=1}^n y_i\right)}\right)}{a}.$$

Proof: The risk function of the estimator λ under the LINEX loss function $L_l\left(\hat{\lambda}, \lambda\right)$ is given by the formula

$$R\left(\hat{\lambda}\right) = \int_0^\infty \left(\exp\left(a\left(\hat{\lambda} - \lambda\right)\right) - a\left(\hat{\lambda} - \lambda\right) - 1 \right) \pi_1\left(\lambda | \underline{x}\right) d\lambda. \tag{22}$$

Using Eq. (13) in Eq. (22), we have

$$R\left(\hat{\lambda}\right) = \int_0^\infty \left(\exp\left(a\left(\hat{\lambda} - \lambda\right)\right) - a\left(\hat{\lambda} - \lambda\right) - 1 \right) \left(\frac{\lambda^{nk-1}\left(\sum_{i=1}^n x_i\right)^{nk} e^{-\lambda \sum_{i=1}^n x_i}}{\Gamma nk} \right) d\lambda$$

$$R(\hat{\lambda}) = \frac{\left(\sum_{i=1}^n x_i\right)^{nk}}{\Gamma nk} \left[\begin{array}{l} \exp\left(a\hat{\lambda}\right) \int_0^\infty \left(\exp\left(-\left(a + \sum_{i=1}^n x_i\right)\lambda\right)\right) \lambda^{nk-1} d\lambda - a\hat{\lambda} \int_0^\infty \left(\exp\left(-\sum_{i=1}^n x_i\right)\lambda\right) \lambda^{nk-1} d\lambda \\ + a\int_0^\infty \left(\exp\left(-\sum_{i=1}^n x_i\right)\lambda\right) \lambda^{nk} d\lambda - \int_0^\infty \left(\exp\left(-\sum_{i=1}^n x_i\right)\lambda\right) \lambda^{nk-1} d\lambda \end{array} \right].$$

On solving the above expression, we get

$$R(\hat{\lambda}) = \frac{\left(\sum_{i=1}^n x_i\right)^{nk} \exp\left(a\hat{\lambda}\right)}{\left(a + \sum_{i=1}^n x_i\right)^{nk}} - a\hat{\lambda} + \frac{a\Gamma(nk+1)}{\Gamma nk \left(\sum_{i=1}^n x_i\right)} - 1.$$

Minimization of the risk with respect to $\hat{\lambda}$ gives us the optimal estimator i.e.,

$$\frac{\partial}{\partial \hat{\lambda}} \left[R(\hat{\lambda}) \right] = 0$$

$$\hat{\lambda}_l = \frac{nk \log\left(1 + \frac{a}{\left(\sum_{i=1}^n x_i\right)}\right)}{a}. \tag{23}$$

5. Estimation of parameters under Quasi prior

In this section, parameter estimation of Erlang distribution is done by using QUASI prior under different loss functions. The procedure for obtaining the Bayesian estimate is available in Section 3. The estimates of parameter are obtained in the following theorems.

Theorem 5.1: Assuming the loss function $L_p(\hat{\lambda}, \lambda)$, the Bayesian estimator of the rate parameter λ, if the shape parameter k is known, is of the form

$$\hat{\lambda}_p = \frac{\sqrt{(nk-d+1)(nk-d+2)}}{\left(\sum_{i=1}^n x_i\right)}.$$

Proof: The risk function of the estimator λ under the precautionary loss function $L_p(\hat{\lambda}, \lambda)$ is given by the formula

$$R(\hat{\lambda}) = \int_0^\infty \frac{(\hat{\lambda} - \lambda)^2}{\hat{\lambda}} \pi_2(\lambda|\underline{x}) d\lambda \tag{24}$$

Using Eq. (17) in Eq. (24), we have

$$R(\hat{\lambda}) = \int_0^\infty \frac{(\hat{\lambda} - \lambda)^2}{\hat{\lambda}} \left(\frac{\lambda^{nk-d} \left(\sum_{i=1}^n x_i\right)^{nk-d+1} e^{-\lambda \sum_{i=1}^n x_i}}{\Gamma(nk-d+1)} \right) d\lambda$$

$$R(\hat{\lambda}) = \frac{\left(\sum_{i=1}^n x_i\right)^{nk-d+1}}{\Gamma(nk-d+1)} \left[\hat{\lambda} \int_0^\infty \lambda^{nk-d} e^{-\lambda \sum_{i=1}^n x_i} d\lambda + \frac{1}{\hat{\lambda}} \int_0^\infty \lambda^{nk-d+2} e^{-\lambda \sum_{i=1}^n x_i} d\lambda - 2 \int_0^\infty \lambda^{nk-d+1} e^{-\lambda \sum_{i=1}^n x_i} d\lambda \right].$$

On solving the above expression, we get

$$R(\hat{\lambda}) = \hat{\lambda} + \frac{1}{\hat{\lambda}} \frac{\Gamma(nk - d + 3)}{\Gamma(nk - d + 1)\left(\sum_{i=1}^{n} x_i\right)^2} - \frac{2\Gamma(nk - d + 2)}{\Gamma(nk - d + 1)\left(\sum_{i=1}^{n} x_i\right)}.$$

Minimization of the risk with respect to $\hat{\lambda}$ gives us the optimal estimator i.e.,

$$\frac{\partial}{\partial \hat{\lambda}}\left[R(\hat{\lambda})\right] = 0$$

$$\hat{\lambda}_p = \frac{\sqrt{(nk - d + 1)(nk - d + 2)}}{\left(\sum_{i=1}^{n} x_i\right)}. \tag{25}$$

Remark: Replacing d = 1 in Eq. (25), the same Bayes estimator is obtained as in Eq. (19).

Theorem 5.2: Assuming the loss function $l_A(\hat{\lambda}, \lambda)$, the Bayesian estimator of the rate parameter λ, if the shape parameter k is known, is of the form

$$\hat{\lambda}_A = \frac{(nk - d + c + 1)}{\left(\sum_{i=1}^{n} x_i\right)}.$$

Proof: The risk function of the estimator λ under the Al-Bayyati's loss function $L_A(\hat{\lambda}, \lambda)$ is given by the formula

$$R(\hat{\lambda}) = \int_0^{\infty} \lambda^c \left(\hat{\lambda} - \lambda\right)^2 \pi_2(\lambda | \underline{x}) d\lambda. \tag{26}$$

By using Eq. (17) in Eq. (26), we have.

$$R(\hat{\lambda}) = \int_0^{\infty} \lambda^c \left(\hat{\lambda} - \lambda\right)^2 \left(\frac{\lambda^{nk-d}\left(\sum_{i=1}^{n} x_i\right)^{nk-d+1} e^{-\lambda \sum_{i=1}^{n} x_i}}{\Gamma(nk - d + 1)}\right) d\lambda$$

$$R(\hat{\lambda}) = \frac{\left(\sum_{i=1}^{n} x_i\right)^{nk-d+1}}{\Gamma(nk - d + 1)}\left[\hat{\lambda}^2 \int_0^{\infty} \lambda^{nk-d+c} e^{-\lambda \sum_{i=1}^{n} x_i} d\lambda + \int_0^{\infty} \lambda^{nk-d+c+2} e^{-\lambda \sum_{i=1}^{n} x_i} d\lambda - 2\hat{\lambda} \int_0^{\infty} \lambda^{nk-d+c+1} e^{-\lambda \sum_{i=1}^{n} x_i} d\lambda\right].$$

On solving the above expression, we get

$$R(\hat{\lambda}) = \frac{\hat{\lambda}^2 \Gamma(nk - d + c + 1)}{\Gamma(nk - d + 1)\left(\sum_{i=1}^{n} x_i\right)^c} + \frac{\Gamma(nk - d + c + 3)}{\Gamma(nk - d + 1)\left(\sum_{i=1}^{n} x_i\right)^{c+2}} - \frac{2\hat{\lambda}\Gamma(nk - d + c + 2)}{\Gamma(nk - d + 1)\left(\sum_{i=1}^{n} x_i\right)^{c+1}}$$

Minimization of the risk with respect to $\hat{\lambda}$ gives us the optimal estimator, i.e., $\frac{\partial}{\partial \hat{\lambda}}\left[R(\hat{\lambda})\right] = 0$

$$\hat{\lambda}_A = \frac{(nk - d + c + 1)}{\left(\sum_{i=1}^{n} x_i\right)}. \tag{27}$$

Remark: Replacing d = 1 in Eq. (27), the same Bayes estimator is obtained as in Eq. (21).

Theorem 5.3: Assuming the loss function $L_l(\hat{\lambda}, \lambda)$, the Bayesian estimator of the rate parameter λ, if the shape parameter k is known, is of the form

$$\hat{\lambda}_l = \frac{(nk - d + 1) \log \left(1 + \frac{a}{\left(\sum_{i=1}^{n} x_i \right)} \right)}{a}.$$

Proof: The risk function of the estimator λ under the LINEX loss function $L_l(\hat{\lambda}, \lambda)$ is given by the formula

$$R(\hat{\lambda}) = \int_{0}^{\infty} \left(\exp \left(a(\hat{\lambda} - \lambda) \right) - a(\hat{\lambda} - \lambda) - 1 \right) \pi_2(\lambda | \underline{x}) d\lambda. \qquad (28)$$

Using Eq. (17) in Eq. (28), we have.

$$R(\hat{\lambda}) = \int_{0}^{\infty} \left(\exp \left(a(\hat{\lambda} - \lambda) \right) - a(\hat{\lambda} - \lambda) - 1 \right) \left(\frac{\lambda^{nk-d} \left(\sum_{i=1}^{n} x_i \right)^{nk-d+1} e^{-\lambda \sum_{i=1}^{n} x_i}}{\Gamma(nk - d + 1)} \right) d\lambda$$

$$R(\hat{\lambda}) = \frac{\left(\sum_{i=1}^{n} x_i \right)^{nk-d+1}}{\Gamma(nk - 2c + 1)} \left[\begin{array}{l} \exp\left(a\hat{\lambda} \right) \int_{0}^{\infty} \left(\exp\left(- \left(a + \sum_{i=1}^{n} x_i \right) \lambda \right) \right) \lambda^{nk-d} d\lambda - a\hat{\lambda} \int_{0}^{\infty} \left(\exp\left(- \sum_{i=1}^{n} x_i \right) \lambda \right) \lambda^{nk-d} d\lambda \\[12pt] + a \int_{0}^{\infty} \left(\exp\left(- \sum_{i=1}^{n} x_i \right) \lambda \right) \lambda^{nk-d+1} d\lambda - \int_{0}^{\infty} \left(\exp\left(- \sum_{i=1}^{n} x_i \right) \lambda \right) \lambda^{nk-d} d\lambda \end{array} \right].$$

On solving the above expression, we get

$$R(\hat{\lambda}) = \frac{\exp\left(a\hat{\lambda} \right) \left(\sum_{i=1}^{n} x_i \right)^{nk-d+1}}{\left(a + \sum_{i=1}^{n} x_i \right)^{nk-d+1}} - a\hat{\lambda} + \frac{a\Gamma(nk - d + 2)}{\Gamma(nk - d + 1) \left(\sum_{i=1}^{n} x_i \right)} - 1.$$

Minimization of the risk with respect to $\hat{\lambda}$ gives us the optimal estimator i.e.,

$$\frac{\partial}{\partial \hat{\lambda}} \left[R(\hat{\lambda}) \right] = 0$$

$$\hat{\lambda}_l = \frac{(nk - d + 1) \log \left(1 + \frac{a}{\left(\sum_{i=1}^{n} x_i \right)} \right)}{a}. \qquad (29)$$

Remark: Replacing d = 1 in Eq. (29), the same Bayes estimator is obtained as in Eq. (23).

6. Entropy estimation of Erlang distribution

The concept of entropy was introduced by Claude. Shannon [42] in the paper "A Mathematical theory of Communication." This concept of Shannon's entropy is the central role of information theory, sometimes referred as measure of uncertainty. Shannon entropy provides an absolute limit on the best possible lossless encoding or compression of any communication, assuming that the communication may be represented as a sequence of independent and identical distributed random variables. Entropy is typically measured in bits, when the log is to the base 2, and nats, when the log is to the base n.

Shannon's definition of entropy, when applied to an information source, can determine the minimum channel capacity required to reliably transmit the source as encoded binary digit. The entropy of a random variable is defined in terms of its probability distribution and can be shown to be a good measure of randomness or uncertainty. For deriving the entropy of probability distributions, we need the following two definitions that are more discussed in Cover et al. [43].

Definition (i): The entropy of the discrete random variable defined on the probability space is given by

$$H_P(f) = - \sum_{i=1}^{n} p(f = a) \log (p(f = a)).$$

It is obvious that $H_P(f) \geq 0$.

Definition (ii): The entropy of the continuous random variable defined on the real line is given by

$$H(f) = E(-\log(x)) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

In this section, entropy estimation of two parameter Erlang distribution is discussed which given as below.

Theorem 6.1: Let $(x_1, x_2, ..., x_n)$ be n positive independent and identically distributed random samples drawn from a population having Erlang density function Eq. (2), then the Shannon's entropy of two parameter Erlang distribution is given by

$$H(f(x; \alpha, \beta)) = - \log \left(\frac{\lambda^k}{(k-1)!} \right) - (k-1)(\psi(k) - \log \lambda) + k.$$

Proof: Shannon's entropy for a continuous random variable is defined as

$$H(f(x; \lambda, k)) = E(-\log(f(x)) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx \qquad (30)$$

Using Eq. (1) in Eq. (30), we have

$$H(f(x; \alpha, \beta)) = E \left\{ -\log \left(\frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x} \right) \right\}$$

$$H(f(x; \alpha, \beta)) = - \log \left(\frac{\lambda^k}{(k-1)!} \right) - (k-1)E(\log(x)) + \lambda E(x). \qquad (31)$$

Now

$$E(\log(x)) = \int_{0}^{\infty} \log(x) f(x) dx$$

$$E(\log(x)) = \frac{\lambda^k}{(k-1)!} \int_{0}^{\infty} \log(x) x^{k-1} e^{-\lambda x} dx$$

$$Put\ \lambda x = t \ \Rightarrow \partial x = \frac{\partial t}{\lambda}; as\ x \rightarrow 0, t \rightarrow 0\ and\ as\ x \rightarrow \infty, t \rightarrow \infty$$

$$E(\log(x)) = \frac{\lambda^{k-1}}{(k-1)!} \int_0^\infty \log\left(\frac{t}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{k-1} e^{-t} dt$$

$$E(\log(x)) = \frac{1}{(k-1)!}\left[\int_0^\infty \log t \, (t)^{k-1} e^{-t} dt - \log\lambda \int_0^\infty (t)^{k-1} e^{-t}\, dt\right]$$

$$E(\log(x)) = \frac{\Gamma'(k)}{\Gamma k} - \log\lambda$$

$$E(\log(x)) = (\psi(k) - \log\lambda). \tag{32}$$

Also

$$E(x) = \int_0^\infty x f(x;\lambda,k) dx$$

$$E(x) = \frac{\lambda^k}{(k-1)!} \int_0^\infty x^k e^{-\lambda x} dx$$

$$E(x) = \frac{\lambda^k}{(k-1)!} \frac{\Gamma(k+1)}{\lambda^{k+1}}$$

$$E(x) = \frac{k}{\lambda}. \tag{33}$$

Substitute the value of Eqs. (32) and (33) in Eq. (31), we have.

$$H(f(x;\alpha,\beta)) = -\log\left(\frac{\lambda^k}{(k-1)!}\right) - (k-1)(\psi(k) - \log\lambda) + k. \tag{34}$$

7. AIC and BIC criterion for Erlang distribution

For model selection the approach of Akaike information criterion (AIC) and Bayesian information criterion (BIC) based on entropy estimation are used. The Akaike information criterion (AIC) was introduced by Hirotsugu Akaike [44] and proposed it as a measure of goodness of fit of an estimated statistical model. It is a measure of the relative quality of a statistical model for a given set of data. It has been found in information theory that it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data.

The AIC is not a test of the model in the sense of hypothesis testing; rather it is a test between models—a tool for model selection. Given a data set, several competing models may be ranked according to their AIC, with the one having the lowest AIC being the best.

The formula for AIC is given by

$$AIC = 2K - 2\log\left(L\left(\hat{\lambda}\right)\right).$$

where K is the number of parameters and $L\left(\hat{\lambda}\right)$ is the maximized value of the likelihood function for the estimated model.

AICC was first introduced by Hurvich and Tsai [45] and its different derivations were proposed by Burnham and Anderson [46]. AICC is AIC with a correction for finite sample sizes and is given by

$$AICC = AIC + \frac{2K(K+1)}{(n-K-1)}.$$

Burnham and Anderson [47] strongly recommended that we should use AICC instead of AIC when the sample size is small or if K is large. Since AICC converges to AIC as the sample size is getting large. Using AIC, instead of AICC, when the sample size is not many times larger than K^2, increases the probability of selecting models that have too many parameters, i.e., of over fitting. The probability of AIC over fitting can be substantial, in some cases.

The Bayesian information criterion (BIC) also known as Schwarz Criterion is used as a substitute for full calculation of the Bayes' factor since it can be calculated without specifying prior distribution. In BIC, the penalty for additional parameters is stronger than that of the AIC.

The formula for the BIC is given by

$$BIC = K \log n - 2 \log \left(L(\hat{\lambda}) \right).$$

where K is the number of parameters, n is the sample size and $L(\hat{\lambda})$ is the maximized value of the likelihood function.

The AIC and BIC of two parameter Erlang distribution are obtained in this section, which are given below.

The Shannon's entropy of two parameter Erlang distribution is given by

$$SH(ED) = \log (k-1)! - \log \lambda^k - (k-1)E(\log x) + \lambda E(x)$$
$$\hat{H}(ED) = \log (k-1)! - \log \lambda^k - (k-1)\log \overline{x} + \lambda \overline{x}. \tag{35}$$

Also

$$ll(x;\lambda,k) = n \log \lambda^k - n \log (k-1)! + (k-1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i - ll\left(x;\hat{\lambda},\hat{k}\right)$$
$$= n \left[\log (k-1)! - \log \lambda^k - (k-1)\log \overline{x} + \lambda \overline{x} \right]. \tag{36}$$

Comparing Eqs. (35) and (36), we have

$$ll\left(x;\hat{\lambda},\hat{k}\right) = -n\hat{H}(ED).$$

The AIC and BIC methodology attempts to find the model that best explains the data with a minimum of their values, we have.

$ll\left(x;\hat{\lambda},\hat{k}\right) = -n\hat{H}(ED)$, then for Erlang family we have

$$AIC = 2K + 2n\hat{H}(ED) \tag{37}$$

or

$$AIC = 2K - 2ll$$

and

$$BIC = K \log n + 2n\hat{H}(ED) \tag{38}$$

or

$$BIC = K \log n - 2ll.$$

8. Simulation study of Erlang distribution

We have generated the data for Erlang distribution of different sample sizes (15, 30 and 60) in R Software for each pairs of (λ, k), where $(k = 1, 2)$ and $(\lambda = 0.5, 1.0)$. The value for the loss parameter $(C_1 = -1, 1)$ and (a = 0.5, 1.0). The values of extension are (C = 0.5, 1.0). The estimates of rate parameter for each method are calculated. The results are presented in the following tables.

N	k	λ	$\hat{\lambda}_{ML}$	$\hat{\lambda}_p$	$\hat{\lambda}_A$		$\hat{\lambda}_l$	
					c = −1	c = 1	a = 0.5	a = 1.0
15	1	0.5	0.15350	0.16179	0.10530	0.18428	0.13350	0.12597
	2	1.0	0.51869	0.54673	0.43100	0.58896	0.47443	0.44430
30	1	0.5	0.03480	0.03238	0.03238	0.03266	0.04191	0.03259
	2	1.0	0.06850	0.04258	0.07049	0.02085	0.03077	0.02171
60	1	0.5	0.03910	0.03320	0.02856	0.03488	0.03110	0.03060
	2	1.0	0.10349	0.09918	0.08981	0.10244	0.09398	0.09200

ML = maximum likelihood estimate, p = precautionary LF, A = Al-Bayyati's LF, l = LINEX LF.

Table 1.
Mean squared error for $\hat{\lambda}$ under Jeffrey's prior.

N	k	λ	d	$\hat{\lambda}_{ML}$	$\hat{\lambda}_p$	$\hat{\lambda}_A$		$\hat{\lambda}_l$	
						c = −1	c = 1	a = 0.5	a = 1.0
15	1	0.5	0.5	0.15350	0.14710	0.11081	0.18979	0.13901	0.13148
			1.0	0.15350	0.16179	0.10530	0.18428	0.13350	0.12597
	2	1.0	0.5	0.51869	0.51528	0.43951	0.59747	0.48294	0.45281
			1.0	0.51869	0.54673	0.43100	0.58896	0.47443	0.44430
30	1	0.5	0.5	0.03480	0.03286	0.02179	0.02206	0.21630	0.02200
			1.0	0.03480	0.03238	0.03238	0.03266	0.04191	0.03259
	2	1.0	0.5	0.06850	0.05828	0.06361	0.01397	0.02389	0.01483
			1.0	0.06850	0.04258	0.07049	0.02085	0.03077	0.02171
60	1	0.5	0.5	0.03910	0.03206	0.03206	0.03534	0.03156	0.03106
			1.0	0.03910	0.03320	0.02856	0.03488	0.03110	0.03060
	2	1.0	0.5	0.10349	0.09641	0.09021	0.10284	0.09439	0.09241
			1.0	0.10349	0.09918	0.08981	0.10244	0.09398	0.09200

ML = maximum likelihood estimate, p = precautionary LF, A = Al-Bayyati's LF, l = LINEX LF.

Table 2.
Mean squared error for $\hat{\lambda}$ under Quasi prior.

9. Comparison of Erlang distribution (ED) with its sub-models

The flexibility and potentiality of the Erlang distribution is compared with its sub models, which is examined by using different criterions like AIC, BIC and AICC with the help of the following illustration.

Illustration I:

We provide the compatibility of the Erlang distribution (ED) with their sub-models; Chi square and exponential distributions. For this purpose, we generated the data set for Erlang distribution of large sample size (i.e., 200) in R Software for each pairs of (λ, k), where $(k = 2)$ and $(\lambda = 2.5)$. The data analysis is given in the following table:

Model	Parameter estimate	Standard error	Measures			
			$-\log l$	AIC	BIC	AICC
Exponential	$\hat{\lambda}$ = 0.5945	0.04203	315.6853	633.3705	636.6689	581.9245
Chi-square	k = 1.090	0.0586	312.3096	625.6346	628.9329	633.3902
Erlang	k = 2		289.9327	581.8654	585.1637	625.6543
	$\hat{\lambda}$ = 1.1890	0.0594				

Table 3.
AIC, BIC and AICC criterion for different sub-models of ED.
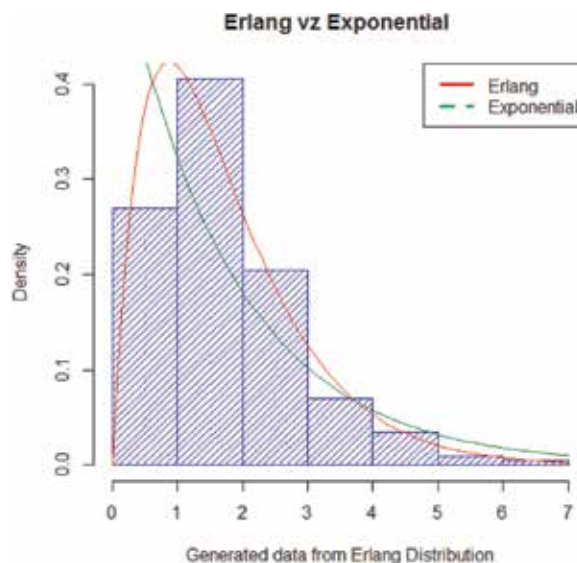
Illustration II:

The data set is taken from Lawless [48]. The observations involves the number of million revolutions between failures for each of 23 ball bearings, the individual bearings were inspected periodically to determine whether "failure" had occurred. Treating the failure times as continuous, the 23 failure times are:

(17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.40, 51.84, 51.96, 54.02, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40)

Model	Parameter estimate	Standard error	Measures			
			$-\log l$	AIC	BIC	AICC
Exponential	$\hat{\lambda}$ = 0.01387	0.002877	121.4324	244.8648	246.0003	245.0248
Chi-square	k = 2.9420	1.1745	177.5835	357.1670	358.3025	357.327
Erlang	k = 1		113.0298	230.0597	232.3307	230.5212
	$\hat{\lambda}$ = 0.05577	0.01683				

Table 4.
AIC, BIC and AICC criterion for different sub-models of ED.

10. Results and discussion

We primarily studied the maximum likelihood (MLH) estimation and Bayesian estimation to estimate the rate parameter of Erlang distribution. In Bayesian method, we use Jeffreys' prior and Quasi prior under three different loss functions. These methods are compared through simulation technique and the results are presented in the **Tables 1** and **2** respectively.

From the results obtained in **Tables 1** and **2**, we observe that in most of the cases, Bayesian estimator under Al-Bayyati's loss function has the smallest mean squared error (MSE) values for Jeffrey's prior and Quasi prior as compared to other loss functions and the maximum likelihood estimator. Thus we can conclude that Bayes estimator under Al-Bayyati's loss function is efficient when the loss parameter **C** is −1.

Also we estimated the unknown parameters of sub-models of Erlang distribution. The Akaike information criterion (AIC), Bayesian information criterion (BIC), and the corrected Akaike information criterion (AICC) are used to compare the candidate distributions. The best distribution corresponds to lower −logL, AIC, BIC, AICC statistics value.

From the results obtained in **Tables 3** and **4**, we observe that the Erlang distribution is a competitive distribution as compared to its sub-models (i.e., exponential distribution and Chi-square distribution). In fact, based on the values of the AIC, BIC and AICC criteria, it shows clear picture that the Erlang distribution provides the best fit for these data among all the models considered.

11. Conclusions

In this paper we have generated three types of data sets with varying sample sizes for Erlang distribution. These data sets were simulated and behavior of the data was checked in case of parameter estimation for Erlang distribution in R Software. By the virtue of the data analysis we are able predict the estimate of rate parameter for Erlang distribution under three different functions by using two different prior distributions. With the help of these results we can also do comparison between loss functions and the priors.

Also the comparison of Erlang distribution with its sub-models was carried out. The results acquired in **Tables 3** and **4**, it shows the clear picture that Erlang distribution performs better as compared to its sub-models. Thus we can say that Erlang distribution is efficient as compared to its sub-models (i.e., exponential distribution and Chi-square distribution) on the basis of the above procedures.

Acknowledgements

The authors acknowledge editor of this journal for encouragement to finalize the chapter. Further, the authors acknowledge profound thanks to anonymous referee for giving critical comments which have immensely improved the presentation of the chapter. Also I extend my sincere thanks to all the authors whose papers I have consulted for this work.

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this chapter.

Author details

Kaisar Ahmad[1]* and Sheikh Parvaiz Ahmad[2]

1 Department of Statistics, Sheikh-ul-Alam Memorial Degree College, Budgam, J&K, India

2 Department of Statistics, University of Kashmir, Srinagar, J&K, India

*Address all correspondence to: ahmadkaisar31@gmail.com

IntechOpen

References

[1] Erlang AK. The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik B. 1909;**20**(6):87-98

[2] Evans M, Hastings N, Peacock B. Statistical Distributions. 3rd ed. New York: John Wiley and Sons, Inc.; 2000

[3] Bhattacharyya SK, Singh NK. Bayesian estimation of the traffic intensity in M/Ek/1 queue Far. East. Journal of Mathematical Sciences. 1994; **2**:57-62

[4] Haq A, Dey S. Bayesian estimation of Erlang distribution under different prior distributions. Journal of Reliability and Statistical Studies. 2001;**4**(1):1-30

[5] Suri PK, Bhushan B, Jolly A. Time estimation for project management life cycles: A simulation approach. International Journal of Computer Science and Network Security. 2009; **9**(5):211-215

[6] Damodaran D, Gopal G, Kapur PK. A Bayesian Erlang software reliability model. Communication in Dependability and Quality Management. 2010;**13**(4):82-90

[7] Jodra P. Computing the asymptotic expansion of the median of the Erlang distribution. Mathematical Modelling and Analysis. 2012;**17**(2):281-292

[8] Zellner A. Bayesian estimation and prediction using asymmetric loss function [PhD thesis]. Journal of American Statistical Association exponential distribution using simulation. Iraq: Baghdad University; 1986. vol. 81, pp. 446-451

[9] Ahmad SP, Ahmad K. Bayesian analysis of Weibull distribution using R Software. Australian Journal of Basic and Applied Sciences. 2013;**7**(9): 156-164. ISSN: 1991-8178

[10] Ahmad K, Ahmad SP, Ahmed A, Reshi JA. Bayesian analysis of generalized gamma distribution using R Software. Journal of Statistics Applications & Probability. 2015;**2**(4): 323-335

[11] Raffia H, Schlaifer R. Applied Statistical Decision Theory. Division of Research, Graduate School of Business Administration, Harvard University; 1961

[12] DeGroot MH. Optimal Statistical Decisions. New York: McGraw-Hill; 1970

[13] Zellner A. Bayesian and non-Bayesian analysis of the log-normal distribution and log normal regression. Journal of the American Statistical Association. 1971;**66**:327-330

[14] Box GEP, Tiao GC. Bayesian Inference in Statistical Analysis. New York: John Wiley and Sons; 1973

[15] Bayes TR. An Essay Towards Solving Problems In The Doctrine Of Chances. Philosophical Transactions of the Royal Society A. 1763;**53**:370-418 Reprinted in Biometrika, 48, 296-315

[16] Laplace PS. Eassi Philosophiquesur Les Probabilities, Paris. This Book Went Through Five Editions (The Fifth was in 1825) Revised By Laplace. The Sixth Edition Appeared In English Translation By Dover Publications, New York, In 1951. While This Philosophical Essay Appeared Separately In 1814, It Also Appeared As A Preface To His Earlier Work. Theoric Analytique Des Probabilites; 1774. pp. 621-656

[17] Jeffrey H. Theory of Probability. III edition (I edition 1939, II edition 1994) ed. Oxford University Press: Clarendon Press; 1961

[18] Anscombe F, Alhnann R. A definition of subjective Probability. The

Annals of Mathematical Statistics. 1963;**34**(1):199-205

[19] Berger JO. Statistical Decision Theory and Bayesian Analysis. 2nd ed. New York: Springer-Verlag; 1985

[20] Berger JO, Bernardo JM. Estimating a product of means: Bayesian analysis with reference priors. American Journal of Mathematical Statistical Association. 1989;**84**:200-207

[21] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. London: Chapman and Hall; 1995

[22] Leonardo T, Hsu JSJ. Bayesian Methods. Cambridge: Cambridge University Press; 1999

[23] De Finetti B. A critical essay on the theory of probability and on the value of science. "probabilismo". Erkenntnis. 1931;**31**:169-223 English translation as "Probabilism"

[24] Bernardo JM, Smith AFM. Bayesian Theory. Chichester, West Sussex: John Wiley and Sons; 2000

[25] Robert CP. The Bayesian Choice. 2nd ed. New York: Springer-Verlag; 2001

[26] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. 2nd ed. Boca Raton, FL: Chapman and Hall/ CRC; 2004

[27] Marin JM, Robert CP. Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer-Verlag; 2007

[28] Carlin BP, Louis TA. Bayesian Methods for Data Analysis. IIIrd ed. Boca Raton, FL: Chapman and Hall/ CRC; 2008

[29] Ibrahim JG, Chen M-H, Sinha D. Bayesian Survival Analysis. New York: Springer Verlag; 2001

[30] Ghosh JK, Delampady M, Samanta T. An Introduction to Bayesian Analysis Theory and Methods. Springer-Verlag; 2006

[31] Bansal AK. Bayesian Parametric Inference. New Dehli: Narosa Publishing House; 2007

[32] Koch KR. Introduction to Bayesian Statistics. New York: Springer-Verlag; 2007

[33] Hoff PD. A First Course in Bayesian Statistical Methods. Springer-Verlag; 2009

[34] Ahmad K, Ahmad SP, Ahmed A. Classical and Bayesian approach in estimation of scale parameter of inverse Weibull distribution. Mathematical Theory and Modeling. 2015;**5**. ISSN 2224-5804

[35] Ahmad K, Ahmad SP, Ahmed A. Classical and Bayesian approach in estimation of scale parameter of Nakagami distribution. Journal of Probability and Statistics. 2016;**2016** Article ID 7581918, 1-8

[36] Jefferys H. An invariant form for the prior probability in estimation problems. Proceedings of The Royal Society Of London, Series. A. 1946;**186**: 453-461

[37] Wald A. Statistical Decision Functions. Wiley; 1950

[38] Norstrom JG. The use of precautionary loss functions in risk analysis. IEEE Transactions on Reliability. 1996;**3**:400-403

[39] Al-Bayyati. Comparing methods of estimating Weibull failure models using simulation [PhD thesis]. Iraq: College of Administration and Economics, Baghdad University; 2002

[40] Klebnov LB. Universal loss function and unbiased estimation. Dokl. Akad.

Nank SSSR Soviet-Maths. Dokl. T. 1972; **203**:1249-1251

[41] Varian HR. A Bayesian approach to real estate assessment. In: Fienberg SE, Zellner A, editors. Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. Amsterdam: North-Holland; 1975. pp. 195-208

[42] Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948;**27**:623-659

[43] Cover TM, Thomas JA. Elements of Information Theory. New York: Wiley; 1991

[44] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. 2nd International Symposium on Information Theory. Budapest: Akademia Kiado; 1973. pp. 267-281

[45] Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika. 1989;**76**:297-307

[46] Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research. 2004; **33**:261-304

[47] Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. Springer-Verlag; 2002

[48] Lawless JF. Statistical Models and Methods for Lifetime Data. 2nd ed. Wiley-InterScience; 2002

Asymptotic Normality of Hill's Estimator under Weak Dependence

Boualam Karima and Berkoun Youcef

Abstract

This note is devoted to the asymptotic normality of Hill's estimator when data are weakly dependent in the sense of Doukhan. The primary results on this setting rely on the observations being strong mixing. This assumption is often the key tool for establishing the asymptotic behavior of this estimator. A number of attempts have been made to relax the assumption of stationarity and mixing. Relaxing this condition, and assuming the weak dependence, we extend the results obtained by Rootzen and Starica. This approach requires less restrictive conditions than the previous results.

Keywords: tail index, Hill's estimator, regularly varying function, linear process, weak dependence

1. Introduction

Extreme value theory (EVT) is a branch of statistics which focus on modeling and measuring extremes events occurring with small probability. Rare events can have severe consequences for human and economic society. The protection against these events is therefore of particular interest. EVT have been extensively applied in various many fields including hydrology, finance, insurance and telecommunications. Unlike most traditional statistical analysis that deal with the center of the underlying distribution, EVT enables us to restrict attention to the behavior of the tails of the distribution which is strongly connected to limiting distribution of extremes values, i.e., maximum or minimum of a sample.

Let $X_1, X_2, ..., X_n$ be *i.i.d* random variables with a common distribution F and let $X_{(n)} \leq ... \leq X_{(1)}$ the order statistics pertaining to $X_1, X_2, ..., X_n$, where $X_{(n)} = \text{Min}(X_i)$ and $X_{(1)} = \text{Max}(X_i)$. Suppose that there exist two normalizing constants a_n, b_n, $(b_n > 0)$ and a nondegenerate distribution function H such that $F^n(b_n x + a_n) \to H(x)$, for every continuity point x of H, then H belongs to one of the three types

- Type I ($\beta > 0$): $\Phi_\beta(x) = \begin{cases} \exp\left(-x^{-\beta}\right) & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$

This class is often called the Frechet class of distributions (fat tailed distribution).

- Type II ($\beta < 0$): $\Psi_\beta(x) = \begin{cases} \exp\left(-\left(-x^\beta\right)\right) & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$

This class is called Weibull class of distributions (short tailed distributions).

- Type III ($\beta = 0$): $\Lambda_0(x) = \exp\left(-e^{-x}\right), \quad \forall x \in \mathbb{R}$,

called the Gumbel type (moderate tail).

This result is known as Fisher-Tippett theorem (see [14]) or the extreme value theorem.

These three family of distributions can be nested into a single representation called the generalized extreme value distribution (GEV) and is given by

$$H_\gamma(x) = \exp\left(-(1 + \gamma x)^{-\frac{1}{\gamma}}\right), \quad 1 + \gamma x > 0$$

This representation is useful in practice since it nets three types of limiting distributions behavior in one framework.

For a positive $\gamma = \frac{1}{\beta}$, we recover the Frechet distribution with, negative $\gamma = \frac{1}{\beta}$ corresponds to the Weibull type, and the limit case $\gamma \to 0$ describes the Gumbel family. The shape parameter γ governs the tail behavior of the distribution. The extreme value theorem remains true if condition of independence of the rv's is replaced by the requirement that the form a stationary sequences satisfying a weak dependence condition called distributional mixing condition (e.g., Leadbetter et al. [20]).

The problem of estimating the tail index has received much attention and a variety of estimators have been proposed in the literature in the context of $i.i.d$ observations, see Hill [16], Pickands [23], Dekkers and De Haan [12]. We focus on the popular Hill estimator (available only for $\beta > 0$) based on the k- upper order statistics and defined as follows.

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}} \tag{1}$$

where $k = k_n$ and $(k_n)_n$ is an intermediate sequence that is, $k_n \to \infty$, $\frac{k_n}{n} \to 0$ as $n \to \infty$ The asymptotic behavior of this estimator has been extensively investigated in the $i.i.d$ setup. Mason [22] proved weak consistency of $H_{k,n}$ for any intermediate sequence k_n and Deheuvels et al. [11] derived strong consistency under the condition that $\frac{k}{log(log(n))} \to \infty$, as $n \to \infty$. Under varying conditions on the sequence k_n and the second-order behavior of F, asymptotic normality of $H_{k,n}$ was discussed among others in Hall [15], Davis and Resnick [7], Csörgo and Mason [5], De Haan and Peng [10], De Haan and Resnick [9].

Hill estimator can still be used for dependent data. In this context, we give below the asymptotic behavior of this estimator.

Hsing [19] and Rootzen et al. [26] established the consistency and the asymptotic normality of $H_{k,n}$ under some general conditions for strictly stationary strong mixing sequences. Brito and Freitas [3] also gave a simplified sufficient condition for consistency, appropriate for applications.

Resnick and Starica [24, 25] proves the weak consistency of Hill's estimator for certain class of stationary sequences with heavy tailed observations which can be approximated by m-dependent sequences. Using this result, they also proved consistency and asymptotic normality of this estimator for an infinite order moving

average and autoregressive sequences with regularly varying marginal distribution. However, Ling and Peng [21] extend their results to an ARMA model with *i.i.d* residuals, based on the estimated residuals, this method can achieve a smaller asymptotic variance than applying hill's estimator to the original data.

Hill [17] proved that $H_{k,n}$ still asymptotically normal for dependent, heterogeneous processes with extremes that form mixingale sequences and for near-Epoch-dependent process.

Hill [18] extends the results of Resnick and Starica [24] and Ling and Peng [21] to a wide range of filtered time series satisfying β- mixing condition. Without using the strong mixing condition, Zhang and McCormick [28], established the asymptotic normality of Hill's estimator, for shot noise sequence provided some mild conditions on the impulse response function.

As mentioned above, the asymptotic normality of Hill's estimator has so far been proved for dependent data under various mixing conditions, but not for weak dependent which is the aim of this note. This notion of weak dependence is more general than the classical frameworks of mixing, associated sequences and Markovian models. This type of dependence covers a broad range of time series models.

In order to establish the asymptotic behavior of Hill's estimator in this setting, we first extend the result of Rootzen et al. [26] to random variables, which fulfill the weak dependence condition. Secondly, we derive the asymptotic normality of the Hill estimator when the observations are generated by a linear process satisfying the η- weak dependence condition (see Boualam and Berkoun [2]). This result extends the work of Resnick and Starica [24].

The novelty of using weak dependence instead of mixing dependence lies in the fact that conditions ensuring the normality of the Hill estimator are weaker than the existing conditions.

To make the chapter self-contained, we present definitions and some important results that we need in the sequel.

2. Definitions and auxiliary results

2.1 Regularly varying functions

We start with some background theory on regular variation.

2.1.1 Regularly varying

A positive measurable function $1 - F$ is called regularly varying function at infinity with index $-\beta$, $\beta > 0$ (written $1 - F \in RV_\beta$) if

$$\lim_{t \to +\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\beta}, \quad x > 0$$

Recall that F belongs to the domain of attraction of $H_{\frac{1}{\beta}}$, $\beta > 0$ if and only if $1 - F(x) \in RV_\beta$.

To establish the asymptotic normality of $H_{k,n}$, a second order regular variation is imposed on the survival function distribution $1 - F$.

2.1.2 Second-order regular condition

A function $1 - F$ is said to be of second-order regular variation with parameter $\rho \leq 0$, if there exists a function $g(t)$ having constant sign with $\lim_{t \to +\infty} g(t) = 0$ and a constant $c \neq 0$ such that

$$\lim_{t \to +\infty} \frac{\frac{1-F(tx)}{1-F(t)} - x^{-\beta}}{g(t)} = cx^{-\beta} \int_1^x \mu^{\rho-1}d\mu, \quad x > 0 \tag{2}$$

Then it is written as $1 - F \in 2RV(-\beta, \rho)$ and $g(t)$ is referred as the auxiliary function of $1 - F$. The convergence in (2) is uniform in x on compact intervals of $(0, +\infty)$.

Under this assumption, de Haan and Peng derived the asymptotic expansion

$$H_{k,n} = \gamma + \frac{\gamma}{\sqrt{k}}Z_k + \frac{g((n/k))}{1-\rho}\left(1 + o_p(1)\right)$$

where $Z_k = \sqrt{k}\left(\frac{\sum_{i=1}^k E_i - 1}{k}\right)$ and E_i is a sequence of i.i.d standard exponential random variables. Hence, choosing k such that $\sqrt{k}g(n/k) = \lambda \neq 0$ leads to asymptotic normality of $\sqrt{k}(H_{k,n} - \gamma)$ with mean $\frac{\lambda}{1-\rho}$ and variance γ^2.

2.2 Strong mixing condition and weak dependence

Several ways of modeling dependence have already been proposed. One of the most popular is the notion of strong mixing introduced by Rosenblatt [27].

2.2.1 Strong mixing

The sequence $(X_n)_n$ is called strongly mixing with mixing coefficient

$$\alpha_{n,l} = \sup\left\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_{1,p}, B \in \mathcal{F}_{p+l,n}, 1 \leq p \leq n - l\right\}$$

if
$$\lim_{n \to \infty} \alpha_{n,l} = 0$$

$\mathcal{F}_{i,j}$ is the $\sigma-$field generated by $\left(X_p : i \leq p \leq j\right)$

It turns out certain classes of processes are not mixing. Inspired by such problems, and in order to generalize mixing and other dependence, Doukhan and Louhichi introduced a new weak dependence condition.

Recall that random variables U, V with values in a measurable space χ are independent if for some rich enough class $\mathcal{F}$ of numerical functions on χ

$$Cov(f(U), g(V)) = 0, \quad \forall f, g \in \mathcal{F}$$

Weakening this assumption leads to definition of weak dependence condition. More precisely, assume that, for convenient functions f and g, $\text{cov}\left(f(\text{"past"}), g(\text{"future"})\right)$ converge to zero as the distance between the "past" and the "future" converge to infinity. Here "past" and "future" refer to the values of the process of interest. This makes explicit the asymptotic dependence between past and future.

Now we describe the notion of weak dependence (in the sense of Doukhan and Louhichi) considered here (see [13]).

2.2.2 Weak dependence

A process $(X_n)_n$ is called $(\varepsilon, \mathcal{E}_n, \Psi)$-weakly dependent if there exists a function $\Psi : \mathbb{R}_+^2 \to \mathbb{R}_+$ and a sequence $\varepsilon = (\varepsilon_l)_l \in \mathbb{N}$ decreasing to zero at infinity, such that for any $(h, k) \in \mathcal{E}_u \times \mathcal{E}_v$ and $(u, v) \in \mathbb{N}^2$

$$Cov\left(h(X_{i_1}, ..., X_{i_u}), k\left(X_{j_1}, ..., X_{j_v}\right)\right) \leq \Psi(Lip(h), Lip(k), u, v)\varepsilon_l$$

For any $(i_1, ..., i_u)$ and $(j_1, ..., j_v)$ with $i_1 < .... < i_u \leq i_u + l \leq j_1 < .... < j_v$.

ε_n denotes the class of real Lipschitz functions, bounded by 1 and defined on $\mathbb{R}^n$ ($n \in \mathbb{N}^*$). $Lipf$ denotes the Lipschitz modulus of continuity of function f, that is

$$Lip(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\|x - y\|_1}$$

with $\|x - y\|_1 = \sum_{i=1}^{n} |x_i - y_i|$.

Specific functions Ψ yield variants of weak dependence appropriate to describe various examples of models:

- η-weakly dependence for which $\Psi(Lip(h), Lip(k), u, v) = uLip(h) + vLip(k)$

- λ-weakly dependence for which
  $\Psi(Lip(h), Lip(k), u, v) = uLip(h) + vLip(k) + uvLip(h)Lip(k)$

- κ-weakly dependence for which $\Psi(Lip(h), Lip(k), u, v) = uvLip(h)Lip(k)$

- ζ-weakly dependence for which $\Psi(Lip(h), Lip(k), u, v) = min(u, v)Lip(h)Lip(k)$

Several class of processes satisfy the weak dependence assumption, as the Bernoulli shift, a Gaussian or an associated process, linear process, $GARCH(p, q)$ and $ARCH(\infty)$ processes (more examples and details can be found in the Dedecker et al. [8]).

The coefficients of weak dependence have some hereditary properties. If the sequence $(X_t)_t$ is κ, λ or θ weakly dependent, then for a Lipschitz function h, the sequence $(h(X_t))_t$ is also weakly dependent.

Mixing conditions refer to σ−algebras rather than to random variables. The main inconvenience of mixing coefficients is the difficulty of checking them. The weak dependence in the sense of Doukhan is measured in terms of covariance which is much easier to compute than mixing coefficients.

3. Asymptotic normality of Hill's estimator under strong mixing condition

In order to proof the asymptotic normality of Hill estimator, we use the approach of Rootzen described in the following.

Let $(Y_n)_n$ be a sequence of stationary strong mixing random variables with mixing coefficients α_{n, l_n} tending to zero at infinity and $l_n = o(n)$. Suppose that the common distribution function F of Y_n is such that

$$\lim_{t \to +\infty} \frac{1 - F(t + x)}{1 - F(t)} = e^{-\beta x}, \quad x \geq 0 \tag{3}$$

i.e., $1 - F(x)$ decays approximately in an exponential manner $e^{-\beta x}$ as $x \to \infty$ or (by log transformations) as an approximate Inverse power law in the sense of regular variation.

Rootzen et al. [26] considered the estimator

$$\beta_n^* = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)} - Y_{(k)} \tag{4}$$

Where $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ are the order statistics pertaining to a sample $Y_1, Y_2, ..., Y_n$.

Under certain conditions, they proved that

$$\sqrt{\frac{k_n}{\lambda_n}} (\beta_n^* - \beta_n) \xrightarrow{d} N(0, 1) \tag{5}$$

where $\beta_n = \frac{n}{k_n} E(Y_1 - u_n)_+$ and $\lambda_n = \frac{n}{k_n r_n} \mathrm{var} \sum_{i=1}^{r_n} \left\{ (Y_j - u_n) 1_{\{Y_j - u_n \geq 0\}} - \frac{1}{\beta} 1_{\{Y_j - u_n \geq 0\}} \right\}$.

The sequences (u_n), (r_n) are chosen such that $\lim_{n \to \infty} p_n \left(\alpha_{n, l_n} + \frac{l_n}{n} \right) = 0$, $\lim_{n \to \infty} \frac{n(1 - F(u_n))}{k_n} = 1$ and

$$r_n = \left[\frac{n}{p_n} \right], \quad \text{with} \quad p_n \to \infty, \quad \frac{r_n}{n} \to 0 \quad \text{as} \quad n \to \infty \tag{6}$$

Note that if we replace $Y = \log X$ in (13), we find the expression of the Hill estimator.

3.1 Hill's estimator in case of infinite order moving average process

Resnick and Starica [24] generalize the Hill estimator for more general settings with possibly dependent data especially for infinite moving average model and AR(p) process.

For a sequence $(X_n)_n$ of random variables generated by a strong mixing linear process, with common distribution F satisfying the following von Mises condition

$$\lim_{t \to +\infty} \frac{tf(t)}{1 - F(t)} = \beta \tag{7}$$

Resnick and Starica [24] have adopted the approach of Rootzen applied to $(Y_n)_n = (logX_n)_n$ for proving the normality of Hill's estimator. It is well known that if (15) holds then $1 - F \in RV_{-\beta}$.

Let $(X_t)_t$ be a strictly stationary linear process defined by

$$X_t = \sum_{i \geq 0} c_i \varepsilon_{t-i} \tag{8}$$

$(\varepsilon_t)_t$ is an *i.i.d* sequence of random variables with marginal distribution satisfying

$$\overline{G}(x) = 1 - G(x) = x^{-\beta} l(x), \quad x > 0, \quad \beta > 0 \tag{9}$$

l is a slowly varying function at infinity and $(c_i)_i$ is a sequence of real numbers satisfying certain mild summability conditions.

Throughout this paper, assume that:

$$\sum_{j=0}^{\infty} |c_j|^{\delta} < \infty, \text{ for some } \quad 0 < \delta < 1 \wedge \beta \tag{10}$$

then (Cline [4]) $\sum_{j=0}^{\infty} c_j \varepsilon_j < \infty$ which implies that $\sum_{j=0}^{\infty} |c_j| \|\varepsilon_j| < \infty$ (Datta and McCormick [6]). Next, assume that

$$\sum_{j=1}^{\infty} \sum_{k=0}^{\infty} |c_{j+k}|^{\beta} \wedge |c_k|^{\beta} \log \left(\frac{|c_{j+k}| \vee |c_k|}{|c_{j+k}| \wedge |c_k|} \right) < \infty \qquad (11)$$

then $X_t = \sum_{i \geq 0} c_i \varepsilon_{t-i}$ is regularly varying.

As a direct consequence of the lemma 2.1 for Resnick and Starica [24] we have

$$\lambda_n \to \lambda = \frac{1}{\beta^2} \left(1 + 2 \frac{\sum_{j=1}^{\infty} \sum_{k=0}^{\infty} |c_k|^{\beta} \wedge |c_{k+j}|^{\beta}}{\sum_{j=0}^{\infty} |c_k|^{\beta}} \right), \text{ as } n \to \infty \qquad (12)$$

Note that λ is finite and depends only on the coefficients c_j.

Theorem 3.1 *(Resnick and Starica [24]) Let $(X_t)_t$ be a strongly mixing linear process and assume that conditions (7), (9), (10) and (11) hold. If the intermediate sequence k is such that*

$$\liminf_{n \to \infty} \frac{n}{k^{\frac{3}{2}}} > 0 \quad or \quad \limsup_{n \to \infty} \frac{n}{k^{\frac{3}{2}}} < \infty \qquad (13)$$

and $1 - F \in 2RV(-\beta, \rho)$ with the auxiliary function g satisfying:

$$\sqrt{n} g \left(b \left(\frac{n}{k} \right) \right) \to 0, \quad as \, n \to \infty \text{ where } b \text{ is the quantile function} \qquad (14)$$

Then

$$\sqrt{k} \left(H_{k,n} - \frac{1}{\beta} \right) \to N(0, \lambda)$$

Note that the second order condition imposed on F implies condition (11) required by Rootzen (see Rootzen et al. [26], Appendix. p44). Condition (13) on the intermediate sequence k allows us to prove the existence of sequence $(r_n)_n$ previously defined.

3.2 Hill's estimator in case of AR(p)process

Similar result to 3.1 where obtained by Rootzen et al. [26] for $AR(p)$ process. Consider a stationary, *pth*-order autoregression $(X_t)_t$ satisfying

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t, \quad t \in \mathbb{N} \qquad (15)$$

We assume that the common distribution of *i.i.d* sequence ε_i satisfy condition (9). Under mild conditions the process (15) has a causal representation of the form (8); if these conditions are not verified then the procedure of applying the Hill estimator directly to an autoregressive process is first to estimate the autoregressive coefficients and then estimating β using estimated residuals.

We assume that we have a sequence $\hat{\phi}^{(n)} = \left(\hat{\phi}_1^{(n)}, ..., \hat{\phi}_p^{(n)} \right), n \geq 1$, of consistent estimators for the coefficients of the autoregression such that $d(n) \left(\hat{\phi}^{(n)} - \phi \right) \to S$ where S is nondegenerate random vector and $d(n) \to \infty$. So that $\varepsilon_t - \hat{\varepsilon}_t^{(n)} = \sum_{i=1}^{p} \left(\hat{\phi}_i^{(n)} - \phi_i \right) X_{t-i}$.

Applying the Hill estimator to the estimated residuals $\left|\hat{\varepsilon}_1^{(n)}\right|, \left|\hat{\varepsilon}_2^{(n)}\right|, ..., \left|\hat{\varepsilon}_n^{(n)}\right|$, Resnick and Starica [24] obtained that, if the distribution $\overline{G}_{|\varepsilon|} \in 2RV(-\beta, \rho)$ and the sequence k is chosen to satisfy the condition (13) and $\frac{\sqrt{k}b\left(\frac{n}{\sqrt{k}}\right)}{b\left(\frac{n}{k}\right)} = o(d(n))$, as $n \to \infty$ then,

$$\sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{\left|\hat{\varepsilon}_{(i)}^{(n)}\right|}{\left|\hat{\varepsilon}_{(k+1)}^{(n)}\right|} - \frac{1}{\beta}\right) \xrightarrow{d} N\left(0, \frac{1}{\beta^2}\right)$$

For AR(p) process, the approach used is quite different than the previous one. Instead, of working with the original observations, the authors used the estimated residuals in order to get the asymptotic normality of Hill's estimator. This method achieves a smaller variance of the Hill estimator than the first one.

4. Asymptotic normality of Hill's estimator under weak dependence

Following the approach of Rootzen et al. [26], we investigate the asymptotic normality of the Hill estimator when the observations are drawn from a causal weakly dependent process in Doukhan sense. In order to check the asymptotic normality of the Hill estimator, we first extend the normality asymptotic of β_n^* defined by (13) for η-weakly dependent random variables. Therefore, applying this to the process $(Y_t)_t$ where $Y_t = \log X_t$, we obtain the desired result.

Let $(Y_n)_n$ be a stationary sequence of random variables η-weakly dependent. We suppose that for each sequences $(p_n)_n$ and $(r_n)_n$ the condition (6) is satisfied and $(l_n)_n$ is such that

$$\lim_{n\to\infty}\frac{l_n}{n} = 0, \ \lim_{n\to\infty}\frac{l_n}{r_n} = 0 \ \text{and} \ \lim_{n\to\infty}\frac{nl_n^{\frac{1}{2}-\mu}}{r_n} = 0, \ \mu > \frac{1}{2} \qquad (16)$$

To establish the asymptotic normality of Hill's estimator, we need to show that under suitable conditions and even if the function *logarithm* does not satisfy the conditions of proposition 2.1 of [8], the sequence $(Y_t)_t = (logX_t)_t$ is η-weakly dependent and possess the hereditary property.

Lemma 4.1 *(Boualam and Berkoun [2]) Let $(X_t)_t$ be a stationary sequence of positive random variables η -weakly dependent. Suppose that there exists a constant $C > 0$, such that $\|X_1\|_p \leq C$, with $p > 1$ then $(Y_t)_t$ where $Y_t = logX_t$ is also η -weakly dependent with $\eta_Y(r) = \mathcal{O}\left(\eta^{\frac{p}{p-1}}(r)\right)$.*

Let $(X_t)_t$ be a causal linear process given by (8) where

$$c_k = \mathcal{O}(|k|^{-\mu}), \ \text{ with } \ \mu > 1/2 \qquad (17)$$

then $(X_t)_t$ is η-weak dependent with $\eta_{l_n} = \mathcal{O}\left(\frac{1}{l_n^{\mu-1/2}}\right)$ (see Bardet et al. [1]).

Now, we extend theorem 4.3 of Rootzen et al. [26] obtained for strong mixing sequences to η-weakly dependent random variables.

Theorem 4.1 *(Boualam and Berkoun [2]) Let $(Y_n)_n$ be a stationary sequence of η-weakly dependent random variables. If condition $\lim_{n\to\infty} p_n\left(\alpha_{n,l_n} + \frac{l_n}{n}\right) = 0$ of theorem 4.3 of Rootzen et al. [26] is replaced by (16), then*

$$\sqrt{\frac{k_n}{\lambda_n}}(\beta_n^* - \beta_n) \xrightarrow{d} N(0,1) \tag{18}$$

The above results allows us to state our main result which extend the result obtained by Resnick and Starica [24] for strong mixing to weak dependent sequences.

Theorem 4.2 *(Boualam and Berkoun [2]) Let* $(X_t)_t$ *be a linear process given by* (16) *with common distribution F, satisfying assumptions* (7), (10), (11), (13), (14), (16) *and* (17) *then*

$$\sqrt{k}\left(H_{k,n} - \frac{1}{\beta}\right) \to N(0,\lambda)$$

5. Conclusion

In a primary work, Hsing showed the asymptotic normality of Hill's estimator in a weak dependent setting under suitable mixing and stationary conditions. Similar results have derived for data with several types of dependence or some specific structures. These conditions have been considerably weakened in Hill. We extend the results obtained by Rootzen and Resnick and Starica. The contribution of this note is threefold. First, the weak dependence in the sense of Doukhan is more general than the framework of mixing and several class of processes possesses this type of dependence. It is important to stress that this dependence allows us to prove the asymptotic normality of the Hill estimator without requiring the assumption that the linear process enjoys the strong mixing property. Consequently, the conditions ensuring the asymptotic normality are weakened with our approach. Second, mixing is hard to verify and requires some regularity conditions. However, using weak dependence which focus on covariances is much easier to compute and this assumption is more often checked by several process. Third, our work can be extended to linear process with dependent innovations (under mild conditions, linear process with dependent innovations is η-weak dependent).

Author details

Boualam Karima and Berkoun Youcef*
Laboratoire de Mathématiques Pures et Appliquées, Faculté des Sciences,
Université Mouloud Mammeri, Tizi Ouzou, Algeria

*Address all correspondence to: youberk@yahoo.com

IntechOpen

References

[1] Bardet JM, Doukhan P, León JF. A functional limit theorem for η-weakly dependent processes and its applications. Statistical Inference for Stochastic Processes. 2008;**11**(3): 265-280

[2] Boualam K, Berkoun Y. Hill's estimator under weak dependence. Communications in Statistics - Theory and Methods. 2017;**46**(18):9218-9229

[3] Brito M, Freitas ACM. Consistent estimation of the tail index for dependent data. Statistics & Probability Letters. 2010;**80**:1835-1843

[4] Cline D. Estimation and linear prediction for regression, autoregression and ARMA with infinite variance data [thesis]. Fort Collins, CO, USA: Deptartment of Statistics, Colorado State University; 1983

[5] Csörgo S, Mason D. Central limit theorems for sums of extreme values. Mathematical Proceedings of the Cambridge Philosophical Society. 1985; **98**:547-558

[6] Datta S, McCormick WP. Inference for the tail parameters of a linear process with heavy tailed innovations. Annals of the Institute of Statistical Mathematics. 1998;**50**(2):337-359

[7] Davis R, Resnick S. Tail estimates motivated by extreme value theory. The Annals of Statistics. 1984;**12**:1467-1487

[8] Dedecker J, Doukhan P, Lang G, León JR, Louhichi S, Prieur C. Weak Dependence: With Examples and Applications. New York: Springer-Verlag; 2007

[9] De Haan L, Resnick SI. On asymptotic normality of the hill estimator. Communications in Statistics. Stochastic Models. 1997;**4**:849-866

[10] De Haan L, Peng L. Comparison of tail index estimators. Statistica Neerlandica. 1998;**52**(1):60-70

[11] Deheuvels P, Haeusler E, Mason DM. Almost sure convergence of the Hill estimator. Mathematical Proceedings of the Cambridge Philosophical Society. 1988;**104**:2, 371-381

[12] Dekkers ALM, De Haan L. On the estimation of the extreme-value index and large quantile estimation. The Annals of Statistics. 1989;**17**:1795-1832

[13] Doukhan P, Louhichi S. A new weak dependence condition and applications to moment inequalities. Stochastic Processes and their Applications. 1999; **84**:313-342

[14] Fisher RA, Tippett LHC. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society. 1928;**24**:180-190

[15] Hall P. On simple estimates of an exponent of regular variation. Journal of the Royal Statistical Society, Series B. 1982;**44**(1):37-42

[16] Hill BM. A simple general approach to inference about the tail of a distribution. The Annals of Statistics. 1975;**3**(5):1163-1174

[17] Hill JB. On tail index estimation for dependent, heterogeneous data. Econometric Theory. 2010;**26**: 1398-1436

[18] Hill JB. Tail index estimation for a filtered dependent time series. Statistica Sinica. 2015;**25**(2):609-629

[19] Hsing T. On tail index estimation using dependent data. The Annals of Statistics. 1991;**19**:1547-1569

[20] Leadbetter MR, Lindgren G, Rootzen H. Extremes and Related Properties of Random Sequences and Processes. New York: Springer Verlag; 1983

[21] Ling S, Peng L. Hill's estimator for the tail index of an ARMA model. Journal of Statistical Planning and Inference. 2004;**123**(2):279-293

[22] Mason DM. Laws of large numbers for sums of extreme values. The Annals of Probability. 1982;**10**(3):754-764

[23] Pickands J. Statistical inference using extreme order statistics. The Annals of Statistics. 1975;**3**:119-131

[24] Resnick S, Starica C. Asymptotic behavior of Hill's estimator for autoregressive data. Communications in Statistics. Stochastic Models. 1997;**13**: 703-723

[25] Resnick S, Starica C. Tail index estimation for dependent data. Annals of Applied Probability. 1998;**8**:1156-1183

[26] Rootzen H, Leadbetter M, De Haan L. Tail and quantile estimation for strongly mixing stationary sequences. Techical report. No. 292. Chapel Hill: Center for Stochastic Processes, Departement of Statistics, University of North Carolina; 1990

[27] Rosenblatt M. A central limit theorem and a strong mixing condition. Proceedings of the National Academy of Sciences of the USA. 1956;**42**(1):43-47

[28] Zhang C, McCormick WP. Asymptotic properties of the tail distribution and Hill's estimator for shot noise sequence. Extremes: Statistical Theory and Applications in Science. 2012;**15**(4):407-435

A Study on the Comparison of the Effectiveness of the Jackknife Method in the Biased Estimators

Nilgün Yıldız

Abstract

In this study, we proposed an alternative biased estimator. The linear regression model might lead to ill-conditioned design matrices because of the multicollinearity and thus result in inadequacy of the ordinary least squares estimator (OLS). Scientists have developed alternative estimation techniques that would eradicate the instability in the estimates. Several biased estimators such as Stein estimator, the ordinary ridge regression (ORR) estimator, the principal components regression (PCR) estimator. Liu developed a Liu estimator (LE) by combining the Stein estimator with the ORR estimator. Since both ORR and LE depend on OLS estimator, multicollinearity affects them both. Therefore, the ORR and LE may give misleading information in the presence of multicollinearity. To overcome this problem, Liu introduced a new estimator, which is based on k and d biasing parameters, the authors worked on developing an estimator that would still have the valuable characteristics of the Liu-type estimator (LTE) but have a smaller bias. We are proposing a modified jackknife Liu-type estimator (MJLTE) that was created by combining the ideas underlying both the LTE and JLTE. Under mean square error matrix criteria, the MJLTE is superior to Liu-type estimator (LTE) and jackknifed Liu-type estimator (JLTE). Finally, a real data example and a Monte Carlo simulation are also given to illustrate theoretical results.

Keywords: jackknifed estimators, jackknifed Liu-type estimator, multicollinearity, MSE, Liu-type estimator

1. Introduction

With regression analysis; Is there a relationship between dependent and independent variables? If there is a relationship, what is the power of this relationship? What is the relationship between variables? Is it possible to predict prospective variables and how should they be estimated? What is the effect of a particular variable or group of variables on other variables or variables in the event that certain conditions are checked? Try to search for answers to questions such as. Linear regression is very important, popular method in statistics. According to Web of Science, the number of publications about linear regression between 2014 and 2018 is given in **Figure 1**.

According to **Figure 1**, the number of studies conducted in 2014 is 12,381, while the number of studies conducted in 2018 is 13,137.

The number of publications about linear regression by document types is given in **Figure 2**.

The most common type of document about linear regression is the article. This is followed by proceeding paper, review, and editorial material.

The number of publications about linear regression by research area is given in **Figure 3**.

The most widely published area related to linear regression is engineering, followed by mathematics, computer science, environmental sciences, ecology and other scientific fields.

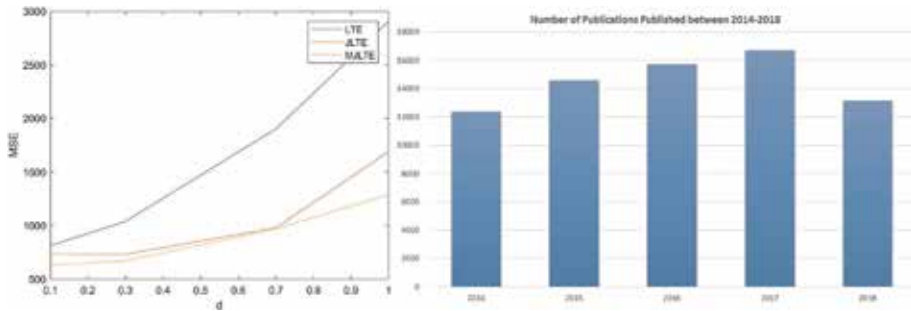The number of publications about linear regression by countries is given in **Figure 4**.

Figure 1.
Number of publications published between 2014 and 2018.
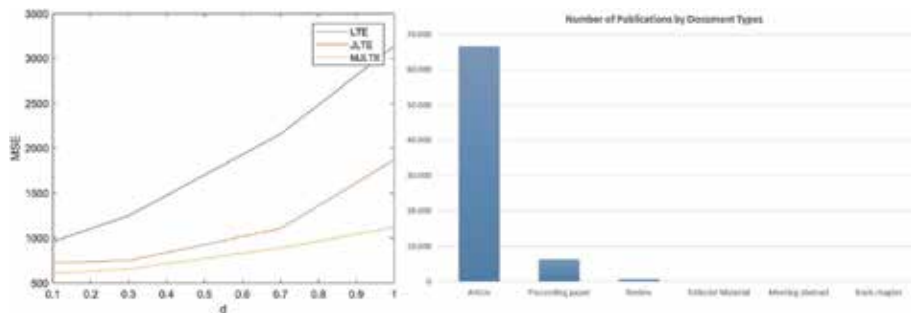
Figure 2.
Number of publications by document types.
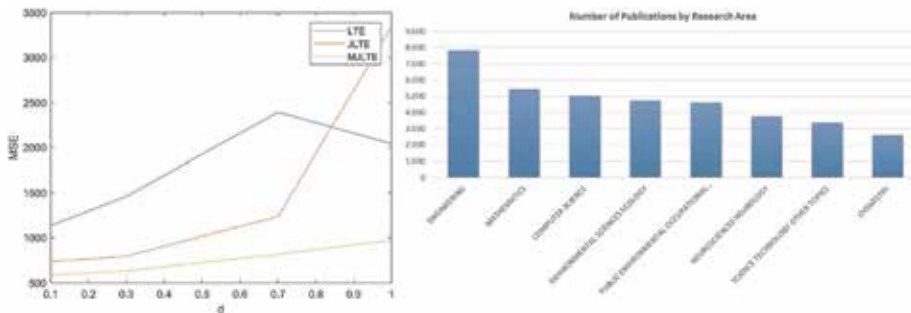
Figure 3.
Number of publications by research area.
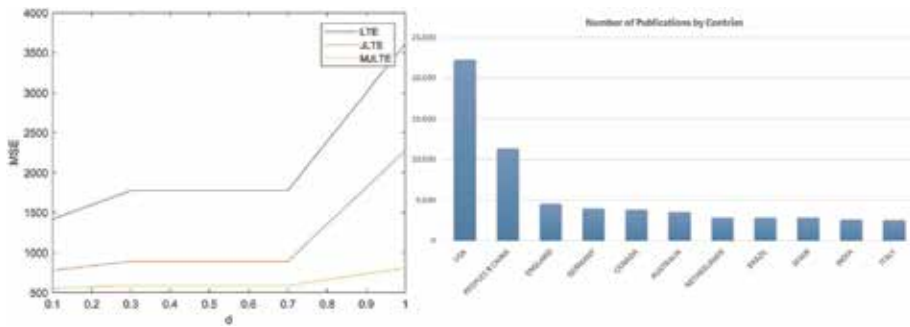
Figure 4.
Number of publications by countries.

The countries with the most publications on linear regression are USA, China, England, Germany, Canada, Australia, respectively.

In regression analysis, the most commonly used method for estimating coefficients is ordinary least squares (OLS). We considered the multiple linear regression model given as

$$y = X\beta + \varepsilon \tag{1}$$

where y is $(n \times 1)$ observable random vector, X is a $(n \times p)$ matrix of non-stochastic (independent) variables of rank p; β is $(p \times 1)$ vector of unknown parameters associated with X, and ε is a $(n \times 1)$ vector of error terms with

$$E(e) = 0, \, Cov(e) = \sigma^2 I \tag{2}$$

In regression analysis, there are several methods to estimate unknown parameters. The most frequently used method is the least squares method (OLS). Apart from this method, there are three general estimation methods: maximum likelihood, generalized least squares, and best linear unbiased estimator BLUE [1].

Since the use of once very popular estimators such as the ordinary least squares (OLS) estimation has become limited due to multicollinearity, which makes them unstable and results in bias and reduced variance of the regression coefficients.

We can give, it is a linear (or close to linear) relationship between the independent variables as the definition of multicollinearity. In the regression analysis, multicollinearity leads to the following problems:

- In the case of multicollinearity, linear regression coefficients are uncertain and the standard errors of these coefficients are infinite.

- The regression coefficients of the multicollinearity increase the variance and covariance of OLS.

- The value of the model R^2 is high but none of the independent variables is significant compared to the partial t test.

- The direction of the related independent variables' relations with the dependent variable may contradict the theoretical and empirical expectations.

- If independent variables are interrelated, some of them may need to be removed from the model. But what variables will be extracted? Removing an incorrect variable from the model will result in a model error. On the other hand, there are no simple rules that we can use to include and subtract the arguments in the model.

Methods for dealing with multicollinearity are collecting additional data, model respecification. Instead of two related variables, the sum of these two variables (as a single variable) can be taken and use of biased estimators. In this book provides information on biased estimators used as OLS alternatives. In literature many researchers have developed biased regression estimators [2, 3].

Examples of such biased estimators are the ordinary ridge regression (ORR) estimator introduced by Hoerl and Kennard [4].

$$\hat{\beta}_k = (X'X + kI)^{-1}X'y \quad k \geq 0 \tag{3}$$

where k is a biasing parameter, in later years researchers combined various estimators to obtain better results. For example, Baye and Parker [5] introduced $(r - k)$ class estimator, which combines the ORR and principal component regression (PCR). In addition, Baye and Parker also showed that $(r - k)$ class estimator is superior to PCR estimator based on the scalar mean square error (SMSE) criterion.

Since both ORR and LE depend on OLS estimator, multicollinearity affects both of them. Therefore, the ORR and LE may give misleading information in the presence of multicollinearity. Liu estimator (LE) was developed by Liu [6] by combining the Stein [7] estimator with the ORR estimator.

$$\hat{\beta}_d = (X'X + I)^{-1}(X'y + d\hat{\beta}) \quad 0 < d < 1 \tag{4}$$

To overcome this problem, Liu [8] introduced a new estimator, which is based on k and d biasing parameters as follows

$$\hat{\beta}_{LTE} = (X'X + k\,I)^{-1}(X'y + d\hat{\beta}) \quad k > 0, \quad -\infty < d < \infty \tag{5}$$

Next, the authors worked on developing an estimator that would still have valuable characteristics of the Liu-type estimator (LTE), but have a smaller bias. In 1956, Quenouille [9] suggested that it is possible to reduce bias by applying a jackknife procedure to a biased estimator.

This procedure enables processing of experimental data to get statistical estimator for unknown parameters. A truncated sample is used calculate specific function of estimators. The advantage of jackknife procedure is that it presents an estimator that has a small bias while still providing beneficial properties of large samples. In this article, we applied the jackknife technique to the LTE. Further, we established the mean squared error superiority of the proposed estimator over both the LTE and the jackknifed Liu-type estimator (JLTE).

The article is organized as follows: The model as well as LTE and the JLTE are described in Section 2. The proposed new estimator is introduced in Section 3. Superiority of the new estimator vis-a-vis the LTE and the JLTE are studied and the performance of the modified Jackknife Liu-type estimator (MJLTE) is compared to that of the JLTE in Section 4. Sections 5 and 6 consider a real data example and a simulation study to justify the superiority of the suggested estimator.

2. The model

We assume that two or more regressors in X are closely linearly related, therefore model suffers from multicollinearity problem. A symmetric matrix $S = X'X$ has an eigenvalue–eigenvector decomposition of the form $S = T\Lambda T'$, where T is an orthogonal matrix and Λ is (real) a diagonal matrix. The diagonal elements of Λ are the eigenvalues of S and the column vectors of T are the eigenvectors of S. The orthogonal version of the standard multiple linear regression models is

$$y = XTT'\beta + \varepsilon = Z\gamma + \varepsilon \tag{6}$$

where $Z = XT$, $\gamma = T'\beta$ and $Z'Z = \Lambda$. The ordinary least squares estimator (OLSE) of γ is given by

$$\hat{\gamma} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y \tag{7}$$

Liu [8] proposed a new biased estimator for γ, called the Liu-type estimators (LTE), and defined as

$$
\begin{aligned}
\hat{\gamma}_{LTE}(k,d) &= (\Lambda + kI)^{-1}(Z'y - d\hat{\gamma}) \quad \text{for } k \geq 0 \quad \text{and} -\infty \leq d \leq +\infty \\
&= (\Lambda + kI)^{-1}\left(Z'y - d\Lambda^{-1}Z'y\right) \\
&= \left[I - (\Lambda + kI)^{-1}(k+d)\right]\hat{\gamma} \\
&= F(k,d)\hat{\gamma}
\end{aligned}
\tag{8}
$$

where

$$F(k,d) = (\Lambda + kI)^{-1}(\Lambda - dI) \tag{10}$$

$\hat{\gamma}_{LTE}$ has bias vector

$$Bias(\hat{\gamma}_{LTE}) = [(F(k,d) - I)\gamma] \tag{11}$$

and covariance matrix

$$Cov(\hat{\gamma}_{LTE}) = \sigma^2 F(k,d)\Lambda^{-1}F(k,d)' \tag{12}$$

By using Hinkley [10], Singh et al. [11], Nyquist [12], and Batah et al. [13] we can propose the jackknifed form of $\hat{\gamma}_{LTE}$. Quenouille [9] and Tukey [14] introduced the jackknife technique to reduce the bias. Hinkley [10] stated that with few exceptions, the jackknife had been applied to balanced models. After some algebraic manipulations, the corresponding jackknife estimator is obtained by deleting the ith observation (z_i', y_i) as

$$(A - z_i z_i')^{-1} = A^{-1} + \frac{A^{-1}z_i z_i' A^{-1}}{1 - z_i' A^{-1} z_i}$$

$$\hat{\gamma}_{LTE_{-i}}(k,d) = (A - z_i z_i')^{-1}(Z'y - z_i y_i) = \left(A^{-1} + \frac{A^{-1}z_i z_i' A^{-1}}{1 - z_i' A^{-1} z_i}\right)(Z'y - z_i y_i)$$

$$= A^{-1}Z'y - A^{-1}z_i y_i + \frac{A^{-1}z_i z_i' A^{-1}}{1 - z_i' A^{-1} z_i}Z'y - \frac{A^{-1}z_i z_i' A^{-1}}{1 - z_i' A^{-1} z_i}z_i y_i$$

$$= \hat{\gamma}_{LTE}(k,d) + A^{-1}z_i y_i \left(1 + \frac{z_i' A^{-1} z_i}{1 - z_i' A^{-1} z_i}\right) + \frac{A^{-1} z_i z_i'}{1 - z_i' A^{-1} z_i} \hat{\gamma}_{LTE}(k,d)$$

$$= \hat{\gamma}_{LTE}(k,d) - A^{-1}z_i \frac{A^{-1} z_i (y_i - z_i' \hat{\gamma}_{LTE}(k,d))}{1 - z_i' A^{-1} z_i} \tag{13}$$

$$= \hat{\gamma}_{LTE}(k,d) - \frac{A^{-1} z_i e_i}{1 - w_i}$$

where z_i' is the $i\,th$ row of Z, $e_i = y_i - z_i' \hat{\gamma}_{LTE}(k,d)$ is the Liu-type residual, $w_i = z_i' A^{-1} z_i$ is the distance factor and $A^{-1} = (\Lambda + kI)^{-1}(I - d\Lambda^{-1}) = F(k,d)\Lambda^{-1}$. In the view of the non-zero value of w_i reflecting the lack of balance in the model, we use the weighted jackknife procedure. Thus, weighted pseudo values are defined as

$$Q_i = \hat{\gamma}_{LTE}(k,d) + n(1 - w_i)\big(\hat{\gamma}_{LTE}(k,d) - \hat{\gamma}_{LTE_{-i}}(k,d)\big)$$

the weighted jackknifed estimator of γ is obtained as

$$\hat{\gamma}_{JLTE}(k,d) = \frac{1}{n}\sum_{i=1}^{n} Q_i = \hat{\gamma}_{LTE}(k,d) + A^{-1}\sum_{i=1}^{n} z_i e_i \tag{14}$$

$$\sum_{i=1}^{n} z_i e_i = \sum_{i=1}^{n} z_i\big(y_i - z_i' \hat{\gamma}_{LTE}(k,d)\big) = (I - A^{-1})Z'y$$

$$\hat{\gamma}_{JLTE}(k,d) = \hat{\gamma}_{LTE}(k,d) + A^{-1}Z'y - A^{-1}\Lambda\, A^{-1}Z'y = (2I - A^{-1}\Lambda)\hat{\gamma}_{LTE}(k,d) \tag{15}$$

However, since $I - A^{-1}\Lambda = I - (\Lambda + k\,I)^{-1}(\Lambda - dI) = I - F(k,d)$, we obtain

$$\hat{\gamma}_{JLTE}(k,d) = (2I - F(k,d))\hat{\gamma}_{LTE}(k,d) \tag{16}$$

From (9) we have

$$\hat{\gamma}_{JLTE}(k,d) = (2I - F(k,d))F(k,d)\hat{\gamma} \tag{17}$$

$$Bias\big(\hat{\gamma}_{JLTE}(k,d)\big) = (I - F(k,d))^2 \gamma \tag{18}$$

Variance of the JLTE as,

$$Cov\big(\hat{\gamma}_{JLTE}(k,d)\big) = \sigma^2(2I - F(k,d))F(k,d)\Lambda^{-1}F(k,d)'(2I - F(k,d))' \tag{19}$$

MSEMs of the JLTE and LTE as

$$\begin{aligned} MSEM\big(\hat{\gamma}_{JLTE}(k,d)\big) &= Cov\big(\hat{\gamma}_{JLTE}(k,d)\big) + Bias\big(\hat{\gamma}_{JLTE}(k,d)\big)\big(\hat{\gamma}_{JLTE}(k,d)\big)' \\ &= \sigma^2(2I - F(k,d))F(k,d)\Lambda^{-1}F(k,d)'(2I - F(k,d))' \\ &\quad + F(k,d)^2 \gamma\gamma'(I - F(k,d))^{2'} \end{aligned} \tag{20}$$

$$MSEM(\hat{\gamma}_{LTE}(k,d)) = \sigma^2 F(k,d)\Lambda^{-1}F(k,d)' + (F(k,d) - I)\beta\beta'(F(k,d) - I) \tag{21}$$

3. Our novel MJLTE estimator

In this section, Yıldız [15] propose a new estimator for γ. The proposed estimator is designated as the modified jackknifed Liu-type estimator (MJLTE) denoted by $\hat{\gamma}_{MJLTE}(k,d)$

$$\hat{\gamma}_{MJLTE}(k,d) = \left[I - (k+d)^2(\Lambda + k\,I)^{-2}\right]\left[I - (k+d)(\Lambda + k\,I)^{-1}\right]\hat{\gamma} \qquad (22)$$

It may be noted that the proposed estimator MJLTE in (22) is obtained as in the case of JLTE but by plugging in the LTE instead of the OLSE. The expressions for bias, covariance and mean squared error matrix (MSEM) of $\hat{\gamma}_{MJLTE}(k,d)$ are obtained as

$$Bias\left(\hat{\gamma}_{MJLTE}(k,d)\right) = -(k+d)(\Lambda + kI)^{-1}W(\Lambda + kI)^{-1}\gamma \qquad (23)$$

$$Cov\left(\hat{\gamma}_{MJLTE}(k,d)\right) = \sigma^2\Phi\Lambda^{-1}\Phi' \qquad (24)$$

$$MSEM\left(\hat{\gamma}_{MJLTE}(k,d)\right) = \sigma^2\Phi\Lambda^{-1}\Phi' + (k+d)^2(\Lambda + kI)^{-1}W(\Lambda + kI)^{-1}$$
$$\gamma\gamma'\left[(\Lambda + kI)^{-1}W(\Lambda + kI)^{-1}\right]' \qquad (25)$$

where $W = I + (k+d)(\Lambda + kI)^{-1} - (k+d)^2(\Lambda + kI)^{-2} = I + F(k,d) - F(k,d)^2$ and $\Phi = (2I - F(k,d))F(k,d)^2$

4. Properties of the MJLTE

One of the most prominent features of our novel MJLTE estimator is that its bias, under some conditions, is less than LTE estimator from which it originates from.

Theorem 4.1. Under the model (1) with the assumptions (2), the inequality $\left\|Bias\left(\hat{\gamma}_{MJLTE}(k,d)\right)\right\|^2 < \left\|Bias(\hat{\gamma}_{LTE}(k,d))\right\|^2$ holds true for $d > 0$ and $k > d$

Proof. From 11 and 23, we can obtain that

$$\left\|Bias\left(\hat{\gamma}_{MJLTE}(k,d)\right)\right\|^2 - \left\|Bias(\hat{\gamma}_{LTE}(k,d))\right\|^2 = (k+d)^2(\Lambda + k\,I)^{-2}\left[W^2 - (\Lambda + k\,I)^2\right]$$
$$(\Lambda + k\,I)^{-2} > 0$$

It is obvious that the difference is greater than 0, because it consists of the product of the squares in the expression above. Thus, the proof is completed.

Corollary 4.1. The bias of the absolute value of the *i th* component of MJLTE is smaller than that of LTE, namely $\left|Bias(\hat{\gamma}_{MJLTE}(k,d)\right|_i < \left|Bias(\hat{\gamma}_{LTE}(k,d)\right|_i$.

Theorem 4.2. *The MJLTE has smaller variance than the LTE*

Proof. From 12 and 24, it can be shown that

$$Cov(\hat{\gamma}_{LTE}(k,d)) - Cov\left(\hat{\gamma}_{MJLTE}(k,d)\right) = \sigma^2 H$$

where

$$H = \left[I + (k+d)(\Lambda + kI)^{-1}\right]\Lambda^{-1}\left[I + (k+d)(\Lambda + kI)^{-1}\right]' - \Phi\Lambda^{-1}\Phi'$$
$$= (I - F(k,d))\left[\Lambda^{-1} - \left(I - F(k,d)^2\Lambda^{-1}(I - F(k,d)^{2'})\right)\right](I - F(k,d))$$

H is a diagonal matrix and ith element

$$h_{ii} = \frac{\left[(\lambda_i + k_i)^4 - (\lambda_i + 2k_i + d_i)^2(\lambda_i - d_i)^2\right](\lambda_i - d_i)^2}{\lambda_i(\lambda_i + k_i)^6}$$

is a positive number. Thus we conclude that H is a positive definite matrix. This completes the proof.

Next, we prove necessary and sufficient condition for the MJLTE to outperform the LTE using the MSEM condition. The proof requires the following lemma.

Lemma 4.1. Let M be a positive definite matrix, namely $M > 0$, α be some vector, then

$M - \alpha\alpha' \geq 0$ if and only if $\alpha'M^{-1}\alpha \leq 1$

Proof. *see Farebrother* [16]

Theorem 4.3. *MJLTE is superior to the LTE in the MSEM sense,* namely$MSEM\ (\hat{\gamma}_{LTE}(k,d)) - MSEM\ (\hat{\gamma}_{MJLTE}(k,d)) > 0$, *if the inequality*

$\Delta_1 = MSEM\ (\hat{\gamma}_{LTE}(k,d)) - MSEM\ (\hat{\gamma}_{MJLTE}(k,d))$ *is nonnegative definite matrix if and if the inequality*

$$\gamma'\left[L^{-1}\big(\sigma^2 H + F^*(k,d)\gamma\gamma'F^*(k,d)'\big)L^{-1'}\right]^{-1}\gamma \leq 1 \qquad (26)$$

is satisfied with $L = F^*(k,d)W$, $F^*(k,d) = F(k,d) - I$ *and* $W = I + F(k,d)$ $-F(k,d)^2$

Proof.

We consider the difference from (21, 25) we have

$$\begin{aligned}\Delta_1 &= MSEM\ (\hat{\gamma}_{LTE}(k,d)) - MSEM\ (\hat{\gamma}_{MJLTE}(k,d))\\ &= \sigma^2 H + F^*(k,d)\gamma\gamma'F^*(k,d)' - L\gamma\gamma'L'\end{aligned} \qquad (27)$$

where

$$\begin{aligned}H &= \left[I + (k+d)(\Lambda + k\,I)^{-1}\right]\Lambda^{-1}\left[I + (k+d)(\Lambda + k\,I)^{-1}\right]' - \Phi\Lambda^{-1}\Phi'\\ &= (I - F(k,d))\left[\Lambda^{-1} - \left(I - F(k,d)^2\Lambda^{-1}(I - F(k,d)^{2'})\right)\right](I - F(k,d))\end{aligned}$$

$W = I + F(k,d) - F(k,d)^2$ is a positive definite matrix. We have seen H is a positive definite matrix from Theorem 2. Therefore, the difference Δ_1 is a nonnegative definite, if and only if $L^{-1}\Delta_1 L^{-1'}$ is a nonnegative definite. The matrix $L^{-1}\Delta_1 L^{-1'}$ can be written as

$$L^{-1}\Delta_1 L^{-1'} = L^{-1}\big(\sigma^2 H + F^*(k,d)\gamma\gamma'F^*(k,d)'\big)L^{-1'} - \gamma\gamma' \qquad (28)$$

Since the matrix $\sigma^2 H + F^*(k,d)\gamma\gamma'F^*(k,d)'$ is symmetric and positive definite, using Lemma 4.1, we may conclude that $L^{-1}\Delta_1 L^{-1'}$ is a nonnegative definite, if and only if the inequality

$$\gamma'\left[L^{-1}\big(\sigma^2 H + F^*(k,d)\gamma\gamma'F^*(k,d)'\big)L^{-1'}\right]^{-1}\gamma \leq 1$$

is satisfied.

4.1 Comparison between the JLTE and the MJLTE

Here, we show that the MJLTE outperforms the JLTE in terms of the sampling variance.

Theorem 4.4. *The variance of* MJLTE *has a smaller variance than that of the* JLTE *for* $d > 0$ *and* $k > d$

Proof. From (19, 24) it can be written as

$$Cov\big(\hat{\gamma}_{JLTE}(k,d)\big) = \sigma^2(2I - F(k,d))F(k,d)\Lambda^{-1}F(k,d)'(2I - F(k,d))'$$
$$= \sigma^2VU\Lambda^{-1}U'V' \qquad (29)$$

and

$$Cov\big(\hat{\gamma}_{MJLTE}(k,d)\big) = \sigma^2\Phi\Lambda^{-1}\Phi' = \sigma^2VUV\Lambda^{-1}V'U'V' \qquad (30)$$

where $V = I - F(k,d)$ and $U = I + F(k,d)$, respectively. It can be shown that

$$Cov\big(\hat{\gamma}_{JLTE}(k,d)\big) - Cov\big(\hat{\gamma}_{MJLTE}(k,d)\big) = \sigma^2\Sigma \qquad (31)$$

where $\Sigma = VU\big(\Lambda^{-1} - V\Lambda^{-1}V'\big)U'V'$, Σ is a diagonal matrix. Then ith the diagonal element of $Cov\big(\hat{\gamma}_{JLTE}(k,d)\big) - Cov\big(\hat{\gamma}_{MJLTE}(k,d)\big)$ is

$$\frac{\sigma^2(\lambda_i + k + 2d_i)^2(\lambda_i - d_i)^2(k + d_i)(2\lambda_i + k + d_i)}{\lambda_i + (\lambda_i + k_i)^6}$$

Hence of $Cov\big(\hat{\gamma}_{JLTE}(k,d)\big) - Cov\big(\hat{\gamma}_{MJLTE}(k,d)\big) > 0$ which completes the proof.

In the following theorem, we have obtained a necessary and sufficient condition for the MJLTE to outperform the JLTE in terms of matrix mean square error. The proof of the theorem is similar to that of Theorem 4.3.

Theorem 4.5.

$\Delta_2 = MSEM\,\big(\hat{\gamma}_{JLTE}(k,d)\big) - MSEM\,\big(\hat{\gamma}_{MJLTE}(k,d)\big)$ *is a nonnegative definite matrix, if and if the inequality*

$$\gamma'\left[L^{-1}\Big(\sigma^2\Sigma + F^*(k,d)^2\gamma\gamma'F^*(k,d)^{2'}\Big)L^{-1'}\right]^{-1}\gamma \le 1 \qquad (32)$$

is satisfied.

Proof. From (20, 25) we have

$$\Delta_2 = MSEM\,\big(\hat{\gamma}_{JLTE}(k,d)\big) - MSEM\,\big(\hat{\gamma}_{MJLTE}(k,d)\big)$$
$$= \sigma^2\Sigma + F^*(k,d)^2\gamma\gamma'F^*(k,d)^{2'} - F^*(k,d)W\gamma\gamma'W'F^*(k,d)'$$

We have seen from Theorem 4.4 that Σ is a positive definite matrix. Therefore, the difference Δ_2 is a nonnegative definite, if and only if $L^{-1}\Delta_2L^{-1'}$ is a nonnegative definite. The matrix $L^{-1}\Delta_2L^{-1'}$ can be written as

$$L^{-1}\Delta_2L^{-1'} = L^{-1}\Big(\sigma^2\Sigma + F^*(k,d)^2\gamma\gamma'F^*(k,d)^{2'}\Big)L^{-1'} - \gamma\gamma'$$

The difference Δ_2 is a nonnegative definite matrix, if and only if $L^{-1}\Delta_2L^{-1'}$ is a nonnegative definite matrix. Since the matrix $\Big(\sigma^2\Sigma + F^*(k,d)^2\gamma\gamma'F^*(k,d)^{2'}\Big)$ is symmetric and positive definite, using Lemma 4.1, we may conclude that $L^{-1}\Delta_2L^{-1'}$ is nonnegative definite, if and only if the inequality

$$\gamma'\left[L^{-1}\Big(\sigma^2\Sigma + F^*(k,d)^2\gamma\gamma'F^*(k,d)^{2'}\Big)L^{-1'}\right]^{-1}\gamma \le 1$$

is satisfied. This confirms our validation. Theorems 4–6 showed that the estimator we proposed was superior to the LTE estimator and JLTE estimator. Accordingly, we can easily say that the MJLTE estimator is better than other estimators LTE, JLTE.

5. Numerical example

To motivate the problem of estimation in the linear regression model, we consider the hedonic prices of housing attributes. The data consists of 92 detached homes in the Ottawa area sold during 1987 (see Yatchew [17]).

Let y be the *sale price (sp)* of the house, X be a 92×9 observation matrix consisting of the variables: *frplc: dummy for* fireplace(s), *grge: dummy for* garage, *lux: dummy for* luxury appointment, *avginc:* average neighborhood income, *dhwy:* distance to highway, *lot area: area of lot, nrbed:* number of bedrooms, *usespc: usable space*. The data are given in **Table 1**.

The eigenvalues of the matrix $X'X$: 9×9 are given by $\lambda_1 = 1.47$, $\lambda_2 = 3.77$, $\lambda_3 = 4.52$, $\lambda_4 = 15.33$, $\lambda_5 = 18.57$, $\lambda_6 = 20.97$, $\lambda_7 = 41.79$, $\lambda_8 = 271.15$ and $\lambda_9 = 239153.68$.

If we use the spectral norm, then the corresponding measure of conditioning of X is the number $\kappa(X) = \sqrt{\lambda_{\max}(X'X)/\lambda_{\min}(X'X)}$ where $\kappa(.) \in [1, \infty)$. We obtained $\kappa(X) = 403.27$, which is large and so X may be considered as being ill-conditioned.

In this case, the regression coefficients become insignificant and therefore, it is hard to make a valid inference or prediction using OLS method. To overcome many of the difficulties associated with the OLS estimates, the LTE. When $\hat{\beta} = (X'X)^{-1}X'y$ and k and d are biasing parameters the use of $\hat{\beta}_{LTE} = (X'X + kI)^{-1}(X'y + d\hat{\beta})$, $k > 0$, $-\infty < d < +\infty$ has become conventional. The LTE estimator will be used for the following example. The original model was used to reconstruct a canonical form as shown in (6) $y = Z\gamma + \varepsilon$. Estimators $\hat{\gamma}_{LTE}, \hat{\gamma}_{JLTE}$ and $\hat{\gamma}_{MJLTE}$ used data $d = 0.10, 0.30, 0.70, 1$ and $k = 0.30, 0.50, 0.70, 1$. Then, the original variable scale was obtained by using the coefficients estimated by these estimators. The individual values of d and k for the scalar MSE (SMSE = trace (MSEM)) of the estimators are shown in **Tables 2–5**. The effects of different values of d on MSE can be seen in **Figures 5–8** that clearly show that the proposed estimator (MJLTE) has smaller estimated MSE values compared to those of the LTE, JLTE.

We observed that for all values of d SMSE(MJLTE) assumed smaller values compared to both SMSE(JLTE) and SMSE(LTE). The estimators' SMSE values are affected by increasing values of k, however the estimator that is affected the least by these changes is our proposed MJLTE estimator. When compared to the other two estimators, the SMSE values of MJLTE gave the best results for both the small and large values of k and d.

6. A simulation study

We want to illustrate the behavior of the proposed parameter estimator by a Monte Carlo simulation. The main purpose of this article is to demonstrate the construction and the details of the simulation which is designed to evaluate the performances of the estimators LTE, JLTE and MJLTE when the regressors are highly intercorrelated. According to Liu [8] and Kibria [18] the explanatory variables and response variable are generated by using the following equations

sellprix	fireplac	garage	luxbath	avginc	crowdist	ncrosdst	disthwy	lotarea	nrbed	usespace	south	west	nsouth	nwest
180	0	1	0	32.3163	0.93428	0	0.63807	3.63297	3	1.23309	0.84	0.409	0	0
135	0	1	0	31.5016	2.18624	0.13942	0.66452	6.5	3	0.84592	1.44	1.645	0.087171	0.16962
165.9	0	1	0	32.0654	2.31148	0.15337	0.43422	5.72	3	0.87508	1.673	1.595	0.12102	0.16276
101	0	0	0	37.8348	2.54381	0.17924	0.26872	3.136	2	0.71445	2.252	1.183	0.20514	0.10622
127	0	0	0	37.8348	2.5458	0.17947	0.16621	2.7	3	0.73789	2.193	1.293	0.19657	0.12131
235	1	0	1	67.0056	2.77147	0.2046	0.22935	6.695	4	1.3518	2.352	1.466	0.21967	0.14505
195	1	1	0	65.8278	3.08747	0.23979	0.30295	4.23	3	1.16829	2.564	1.72	0.25047	0.17991
184.5	1	0	0	62.3053	3.4844	0.28399	0.33009	4.224	3	0.98228	2.785	2.094	0.28258	0.23123
106	1	1	0	38.4946	3.83086	0.32258	0.87056	3.234	2	0.79507	2.148	3.172	0.19003	0.37917
156	1	1	0	52.3552	3.85306	0.32505	0.39006	5	3	0.90239	2.563	2.877	0.25033	0.33869
195	1	0	0	52.3552	3.88283	0.32836	0.67211	4.8	4	1.29465	2.363	3.081	0.22127	0.36668
206	1	1	0	52.3552	3.92493	0.33305	0.69495	4.75	4	1.7111	2.38	3.121	0.22374	0.37217
157	0	1	0	52.3552	3.95888	0.33683	0.43005	5	3	0.901	2.621	2.967	0.25875	0.35104
180	1	0	0	79.4583	3.96236	0.33722	0.12078	4	3	0.95527	3.015	2.571	0.316	0.29669
193	1	1	0	79.4583	3.9626	0.33725	0.19503	6.35545	3	1.50436	3.063	2.514	0.32297	0.28887
230	1	1	0	52.3552	3.97284	0.33839	0.39307	5	3	1.08421	2.661	2.95	0.26456	0.3487
212	0	1	1	53.8647	4.04329	0.34623	0.22395	5	4	1.08724	2.845	2.873	0.2913	0.33814
102	1	0	0	59.5774	4.11494	0.35421	1.18386	4	2	1.24813	2.113	3.531	0.18495	0.42843
137	1	1	0	59.5774	4.13141	0.35605	1.01187	2.4	3	0.753	2.285	3.442	0.20994	0.41622
187	0	1	0	59.5774	4.13723	0.35669	1.00413	6.10686	3	0.91936	2.297	3.441	0.21168	0.41608
103	0	0	0	59.5774	4.16338	0.35961	1.14419	4	2	0.68893	2.193	3.539	0.19657	0.42953
100	0	0	0	59.5774	4.22521	0.36649	1.22908	4	3	1.07483	2.169	3.626	0.19308	0.44147

sellprix	fireplac	garage	luxbath	avginc	crowdist	ncrosdst	disthwy	lotarea	nrbed	usespace	south	west	nsouth	nwest
152	1	1	0	39.7652	4.29688	0.37447	1.02071	9.9	4	1.01668	3.793	2.019	0.42903	0.22094
127	1	1	0	39.5229	4.52879	0.4003	0.6903	5.16	4	1.64743	2.897	3.481	0.29885	0.42157
119.5	0	1	0	35.6	4.59649	0.40784	1.39392	4.9	2	0.9768	4.206	1.854	0.48903	0.1983
103	0	0	0	39.5229	4.63841	0.41251	0.64774	4.6158	2	1.14358	3.023	3.518	0.31716	0.42665
99	0	1	0	30.6514	4.69599	0.41892	1.00164	5	4	0.8655	4.112	2.268	0.47537	0.25511
75	0	0	0	39.5	4.69941	0.4193	0.49382	1.891	3	1.1436	3.192	3.449	0.34171	0.41718
128	1	1	0	39.5229	4.73353	0.4231	0.7881	5	3	1.03911	2.992	3.668	0.31265	0.44723
132	1	0	0	35.6	4.75938	0.42598	1.23504	5.0892	3	1.32579	4.273	2.096	0.49877	0.23151
132	0	0	0	38.1216	4.80701	0.43128	0.35373	6.5415	4	1.25312	3.851	2.877	0.43745	0.33869
134	0	0	0	39.5	5.04763	0.45808	0.27429	4.725	3	0.80536	3.64	3.497	0.4068	0.42377
120	1	1	0	39.8732	5.08006	0.46169	0.58519	5	3	1.2452	4.208	2.846	0.48932	0.33443
125	0	1	0	39.8732	5.21855	0.47712	0.7121	6.96	2	0.62472	4.391	2.82	0.51591	0.33086
135	1	1	0	25.9545	5.23599	0.47906	1.01716	5.4	3	0.88273	4.563	2.568	0.5409	0.29628
139	1	1	0	38.24	5.26572	0.48237	0.92231	4.08	4	1.44414	3.335	4.075	0.36249	0.50309
151	0	1	0	34.4049	5.34316	0.49099	0.96285	6.6	3	1.03205	3.368	4.148	0.36728	0.51311
116.5	0	1	0	38.1216	5.53704	0.51258	0.51961	5.3	3	0.99325	4.544	3.164	0.53814	0.37807
137	1	1	0	50.0548	5.56767	0.516	0.70637	6.6	3	0.8773	3.757	4.109	0.4238	0.50775
149	1	1	0	55.8667	5.75411	0.53676	1.38296	5.6	3	1.19255	3.37	4.664	0.36757	0.58392
167	1	1	0	58.4763	5.8055	0.54248	1.20341	6.4152	4	1.48624	3.565	4.582	0.3959	0.57266
163	1	1	0	58.4763	5.8213	0.54424	1.03375	4.92	4	1.60065	3.716	4.481	0.41784	0.5588
147.5	0	1	0	58.4763	6.0749	0.57248	0.73326	5.1825	3	0.972	4.16	4.427	0.48235	0.55139
237.7	1	1	0	58.4763	6.0752	0.57252	0.62363	5.4495	4	1.99967	4.241	4.35	0.49412	0.54083

sellprix	fireplac	garage	luxbath	avginc	crowdist	ncrosdst	disthwy	lotarea	nrbed	usespace	south	west	nsouth	nwest
168	1	1	0	44.2475	6.2599	0.59309	1.49998	5	3	1.30275	3.706	5.045	0.41639	0.6362
180	1	1	0	58.4763	6.3173	0.59948	1.03263	5.3014	4	1.59782	4.135	4.776	0.47872	0.59929
156	1	0	0	57.7446	6.6398	0.63539	1.09965	5	3	1.41006	4.354	5.013	0.51053	0.63181
145	1	1	0	44.2475	6.6469	0.63618	1.67399	6.572	4	1.22169	3.891	5.389	0.44327	0.68341
144	1	1	0	57.6861	6.6854	0.64047	0.76067	6.825	4	1.689	5.616	3.627	0.69388	0.44161
140	1	1	0	57.6861	6.69752	0.64182	0.63376	6.3	3	1.14778	5.556	3.74	0.68517	0.45712
236	1	1	1	44.2475	6.7398	0.64653	1.59621	4.725	4	1.63184	4.037	5.397	0.46448	0.68451
172	1	1	0	53.5907	6.8231	0.65581	0.98439	5	4	1.22956	4.596	5.043	0.54569	0.63593
148	1	0	0	71.0269	7.1147	0.68828	1.42051	6	3	1.14276	4.501	5.51	0.53189	0.70001
153.5	1	1	0	71.0269	7.1224	0.68914	1.54559	7.3832	4	1.433	4.406	5.596	0.51809	0.71182
154	1	1	0	71.0269	7.157	0.69299	1.21892	5.292	3	1.44025	4.696	5.401	0.56022	0.68506
145.5	1	1	0	50.7101	7.3023	0.70917	0.6233	6.735	4	1.49297	6.044	4.098	0.75607	0.50624
149	1	1	0	50.7101	7.35272	0.71479	0.4738	6.0888	4	1.26687	6	4.25	0.74967	0.5271
138	1	0	0	50.7101	7.3538	0.71491	0.58063	5.3352	4	1.50339	6.062	4.163	0.75868	0.51516
141.5	1	0	0	52.6378	7.44672	0.72525	1.07227	6.54	4	1.03462	6.403	3.802	0.80822	0.46562
125	1	0	0	52.6378	7.46892	0.72773	0.9696	7.02	4	1.16772	6.368	3.903	0.80314	0.47948
130	0	1	0	52.6378	7.52266	0.73371	0.93924	6.5	4	1.09757	6.396	3.96	0.80721	0.48731
132	0	1	0	52.6378	7.55757	0.7376	0.91361	5.1	4	1.20377	6.411	4.002	0.80939	0.49307
132.9	1	1	0	52.6378	7.58783	0.74097	0.88591	6	3	1.06059	6.421	4.043	0.81084	0.4987
122	1	1	0	52.6378	7.6221	0.74479	0.85984	5	3	0.80494	6.435	4.085	0.81287	0.50446
162	1	1	0	51.5087	7.63628	0.74637	0.35916	4.992	4	1.4318	5.705	5.076	0.70681	0.64046
127.5	1	0	0	73.4464	7.6503	0.74793	0.68757	6.785	3	0.84692	5.488	5.33	0.67529	0.67531

sellprix	fireplac	garage	luxbath	avginc	crowdist	ncrosdst	disthwy	lotarea	nrbed	usespace	south	west	nsouth	nwest
87	0	0	0	42.3138	7.6785	0.75107	1.67919	4.6	4	0.74017	4.769	6.018	0.57083	0.76973
139.9	1	1	0	40.9246	7.90353	0.77613	0.40138	5.5	3	1.19235	6.366	4.684	0.80285	0.58666
240	1	1	1	41.95	7.9194	0.77789	1.16065	5.004	4	1.3972	5.316	5.87	0.6503	0.74942
134	1	1	1	52.6378	7.92504	0.77852	1.08098	4.2581	3	1.2298	6.776	4.11	0.86241	0.50789
136.5	1	0	0	42.3138	7.9381	0.77998	1.58184	4.625	3	1.06694	5.008	6.159	0.60555	0.78908
143.5	1	1	0	51.5087	7.98552	0.78526	0.19433	5	4	1.44972	6.038	5.226	0.75519	0.66104
140	1	1	0	57.5706	7.99308	0.7861	0.94096	7.856	3	1.03167	6.743	4.292	0.85762	0.53287
123	1	0	0	51.5087	8.02072	0.78918	0.7685	6.757	4	1.28782	5.665	5.678	0.701	0.72307
147	1	1	0	57.5706	8.06875	0.79453	0.5518	5.04	3	0.96054	6.565	4.691	0.83176	0.58762
134.9	1	1	0	57.5706	8.16678	0.80544	0.77684	5.185	4	1.35781	6.763	4.578	0.86053	0.57211
154	1	1	0	57.5706	8.23868	0.81345	0.85344	5.27	4	1.60938	6.855	4.57	0.87389	0.57102
143.9	1	0	0	40.9246	8.56769	0.85009	0.53129	5	3	0.91112	6.877	5.11	0.87709	0.64512
126	1	1	0	53.8029	8.62504	0.85648	0.48411	5	4	0.88838	6.885	5.195	0.87825	0.65679
118.5	1	0	0	40.0618	8.9486	0.89251	0.10317	5.1	3	1.05885	6.85	5.758	0.87317	0.73405
158	1	1	1	39.5262	8.9552	0.89325	1.86432	5.035	3	1.1192	5.425	7.125	0.66613	0.92164
118	0	0	0	37.1457	9.2964	0.93124	0.18611	5.035	3	0.93595	7.127	5.969	0.91341	0.763
109.25	0	0	0	30.9704	9.4572	0.94915	0.70329	5.4	3	0.67449	7.56	5.682	0.97632	0.72362
124	0	0	0	49.0297	9.5445	0.95887	0.43645	6	3	1.46683	6.863	6.633	0.87505	0.85412
137	1	0	0	35.5188	9.5795	0.96277	1.7827	5	3	0.97388	5.882	7.561	0.73253	0.98147
142	1	1	0	30.5844	9.6273	0.96809	1.59895	5	4	1.4761	6.056	7.484	0.75781	0.97091
120.5	0	0	0	26.9947	9.6483	0.97043	1.51784	5	3	0.69971	6.132	7.449	0.76885	0.9661
123	0	1	0	51.1569	9.7401	0.98066	1.79905	5	2	0.68574	5.97	7.696	0.74531	1

sellprix	fireplac	garage	luxbath	avginc	crowdist	ncrosdst	disthwy	lotarea	nrbed	usespace	south	west	nsouth	nwest
157.5	1	0	0	51.1569	9.744	0.98109	1.71459	5	4	1.07226	6.039	7.647	0.75534	0.99328
115	1	0	0	47.8688	9.7586	0.98272	0.92055	5	3	0.90163	6.651	7.141	0.84425	0.92384
126.5	1	1	0	55.2901	9.8628	0.99432	0.412	5	3	0.902	7.637	6.241	0.98751	0.80033
155	1	1	0	55.2901	9.9138	1	0.49461	5	3	1.41319	7.723	6.216	1	0.7969

Table 1.
Data set.

	d = 0.10	d = 0.30	d = 0.70	d = 1
MSE(LTE)	810.4511	1037.6454	1900.68971	2905.5467
MSE(JLTE)	733.5563	729.5050	977.1382	1688.9649
MSE(MJLTE)	631.2267	669.0754	967.7905	1289.1137

Table 2.
The estimated MSE values of LTE, JLTE and MJLTE k = 0.30.

	d = 0.10	d = 0.30	d = 0.70	d = 1
MSE(LTE)	957.6623	1245.7243	2157.4693	3134.9466
MSE(JLTE)	725.6311	752.2125	1102.8970	1872.6471
MSE(MJLTE)	608.2459	656.6023	892.2214	1115.3394

Table 3
The estimated MSE values of LTE, JLTE and MJLTE k = 0.50.

	d = 0.10	d = 0.30	d = 0.70	d = 1
MSE(LTE)	1133.2567	1459.7360	2393.9079	2042.7127
MSE(JLTE)	734.5155	795.8311	1234.9720	3340.5986
MSE(MJLTE)	587.0096	633.9972	815.8143	973.5845

Table 4.
The estimated MSE values of LTE, JLTE and MJLTE k = 0.70.

	d = 0.10	d = 0.30	d = 0.70	d = 1
MSE(LTE)	1415.148	1774.1222	1774.1222	3613.8006
MSE(JLTE)	779.0405	891.0250	891.0250	2274.5162
MSE(MJLTE)	551.0494	588.8484	588.8484	807.4456

Table 5.
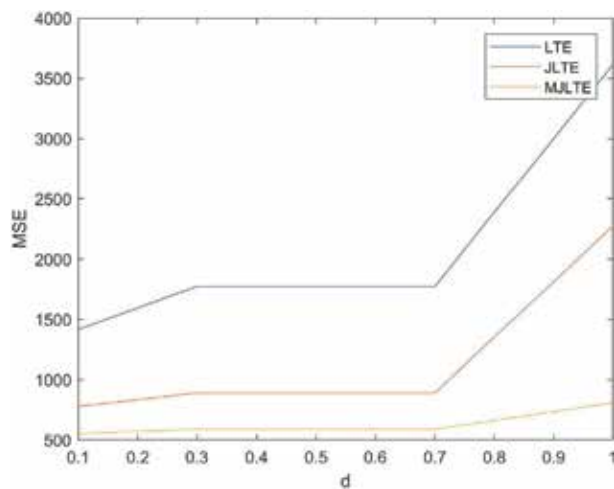*The estimated MSE values of LTE, JLTE and MJLTE k = 1.*

Figure 5.
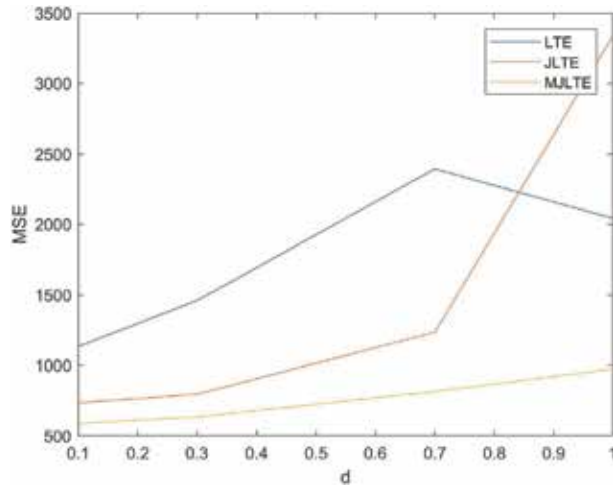*Various MSE of the proposed estimator compared to others for different values of d when k = 0.30.*

Figure 6.
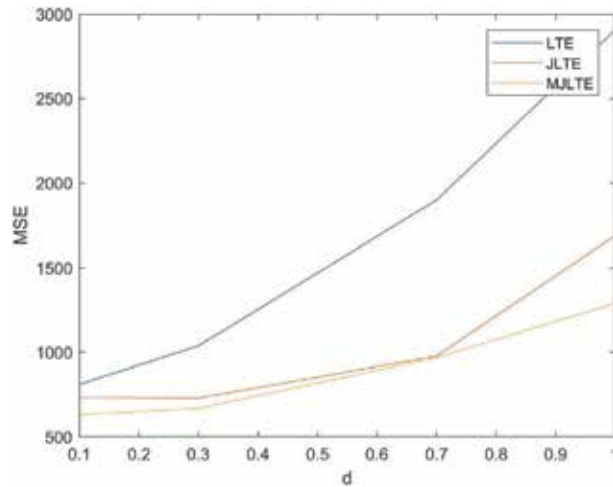*Various MSE of the proposed estimator compared to others for different values of d when k = 0.50.*

Figure 7.
Various MSE of the proposed estimator compared to others for different values of d when k = 0.70.

$$x_{ij} = \left(1 - \gamma^2\right)^{1/2} z_{ij} + \gamma z_{ip}, \ y_i = \left(1 - \gamma^2\right)^{1/2} z_i + \gamma z_{ip} \ i = 1, 2, ..., n, \ j = 1, 2, ..., p$$

where z_{ij} is an independent standard normal pseudo-random number and p is specified so that correlation between any two explanatory variables is given by γ^2. In this study, we used $\gamma = 0.90, 0.95, 0.99$ to investigate the effects of different degrees of collinearity with sample sizes $n = 20, 50$ and 100, while four different combinations for (k, d) are taken as $(0.8, 0.5)$, $(1, 0.7)$, $(1.5, 0.9)$, $(2, 1)$. The standard deviations considered in the simulation study are $\sigma = 0.1; 1.0; 10$. For each choice of γ, σ^2 and n, the experiment was replicated 1000 times by generating new error terms. The average SMSE was computed using the following formula

$$SMSE\left(\hat{\beta}\right) = \frac{1}{1000} \sum_{j=1}^{1000} \left(\beta_j - \beta\right)' \left(\beta_j - \beta\right)$$

Let us consider the LTE, JLTE and MJLTE and compute their respective estimated MSE values with the different levels of multicollinearity. According to the
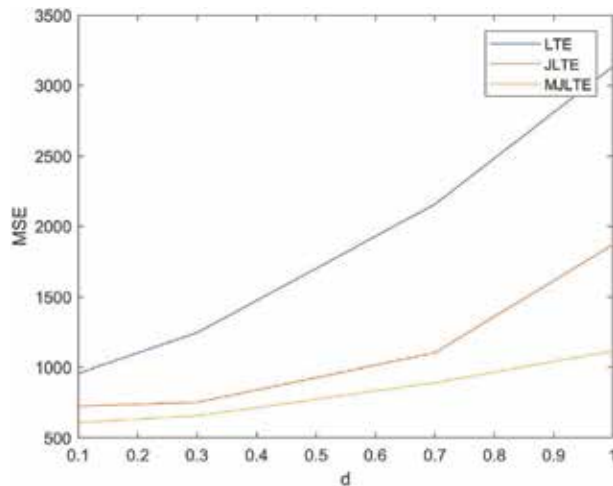
Figure 8.
Various MSE of the proposed estimator compared to others for different values of d when k = 1.

simulation results shown in **Tables 4** and **5** for LTE, JLTE and MJLTE with increasing levels of multicollinearity there was a general increase in the estimated MSE values Moreover, increasing level of multicollinearity also lead to the increase in the MSE estimators for fixed d and k.

In **Table 4**, the MSE values of the estimators corresponding to different values of d are given for $k = 0.70$. For all values of d, the smallest MSE value appears to belong to the MJLTE estimator. The least affected by multicollinearity is MJLTE according to MSE criteria.

In **Table 5**, the MSE values of the estimators corresponding to different values of d are given for $k = 1$.

For all values of d, the smallest MSE value appears to belong to the MJLTE estimator. The least affected by multicollinearity is MJLTE according to MSE criteria.

We can see that MJLTE is much better than the competing estimator when the explanatory variables are severely collinear. Moreover, we can see that for all cases of LTE, JLTE and MJLTE in MSE criterion the MJLTE has smaller estimated MSE values than those of the LTE and JLTE.

7. Conclusion

In this paper, we combined the LTE and JLTE estimators to introduce a new estimator, which we called MJLTE. Combining the underlying criteria of LTE and JLTE estimators enabled us to create a new estimator for regression coefficients of a linear regression model that is affected by multicollinearity. Moreover, the use of jackknife procedure enabled as to produce an estimator with a smaller bias. We compared our MJLTE to its originators LTE and JLTE in terms of MSEM and found that MJLTE has a smaller variance compared to both LTE and JLTE. Thus, MJLTE is superior to both LTE and JLTE under certain conditions.

Author details

Nilgün Yıldız
The Department of Mathematics, Faculty of Arts and Sciences, Marmara University Kuyubaşı, Istanbul, Turkey

*Address all correspondence to: ncelebiyil@gmail.com

IntechOpen

References

[1] Rao RC, Toutenburg H. Linear Models Least Squares and Alternatives. New york: Springer University Press; 1995

[2] Montgomery DC, Peck EA, Vınıng GG. Introduction to Linear Regression Analysis. New york: Wiley; 2006

[3] Chatterjee S, Hadi AS. Regression Analysis by Example. New york: Wiley; 2006

[4] Hoerl A, Kennard R. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;**12**:55-67

[5] Parker DF, Baye MR. Combining ridge and principal component regression: A money demand illustration. Communication in Statistics: Theory and Methods. 1984;**13**(2):197-205

[6] Liu K. A new class of biased estimate in linear regression. Communication in Statistics: Theory and Methods. 1993;**22**(2):393-402

[7] Stein C. Inadmissibility of the usual estimator for mean of multivariate normal distribution. In: Neyman J, editor. Proceedings of the Third Berkley Symposium on Mathematical and Statistics Probability. 1956. pp. 197-206

[8] Liu K. Using Liu-type estimator to combat collinearity. Communications in Statistics. 2003;**32**(5):1009-1020

[9] Quenouille MH. Notes on bias in estimation. Biometrika. 1956;**43**:353-360

[10] Hinkley DV. Jackknifing in unbalanced situations. Technometrics. 1977;**19**:285-292

[11] Singh B, Chaubey YP, Dwivedi TD. An almost unbiased ridge estimator. Sankhya. 1986;**48**:342-346

[12] Nyquist H. Applications of the jackknife procedure in ridge regression. Computational Statistics & Data Analysis. 1988;**6**:177-183

[13] Batah F, Ramanathan TK, Gore SD. The efficiency of modified jackknife and ridge type regression estimators: A comparison. Surveys in Mathematics and its Applications. 2008;**3**:111-122

[14] Tukey JW. Bias and confidence in not quite large samples (abstract). Annals of Mathematical Statistics. 1958;**29**:614

[15] Yıldız N. On the performance of the jackknified Liu-type estimator in linear regression model. Communication in Statistics: Theory and Methods. 2018;**47**(9):2278-2290

[16] Farebrother RW. Further results on the mean square error of ridge regression. Journal of the Royal Statistical Society, Series B. 1976;**38**(B):248-250

[17] Yatchew A. Semiparametric Regression for the Applied Econometrician. Cambridge: Cambridge University Press; 2003

[18] Kibria BMG. Performance of some new ridge regression estimators. Communication in Statistics: Simulation and Computation. 2003;**32**:2389-2413

Edited by Jan Peter Hessling

Statistical practices have recently been questioned by numerous independent authors, to the extent that a significant fraction of accepted research findings can be questioned. This suggests that statistical methodologies may have gone too far into an engineering practice, with minimal concern for their foundation, interpretation, assumptions, and limitations, which may be jeopardized in the current context. Disguised by overwhelming data sets, advanced processing, and stunning presentations, the basic approach is often intractable to anyone but the analyst. The hierarchical nature of statistical inference, exemplified by Bayesian aggregation of prior and derived knowledge, may also be challenging. Conceptual simplified studies of the kind presented in this book could therefore provide valuable guidance when developing statistical methodologies, but also applying state of the art with greater confidence.

IntechOpen