Density Approximant
Based on Noise Multiplied Data

Yan-Xia Lin
National Institute for Applied Statistics Research Australia,
School of Mathematics and Applied Statistics,
University of Wollongong, Australia

Abstract. Using noise multiplied data to protect confidential data has recently drawn some attention. Understanding the probability property of the underlying confidential data based on their masked data is of interest in confidential data analysis. This paper proposes the approach of sample-moment-based density approximant based on noise multiplied data and provides a new manner for approximating the density function of the underlying confidential data without accessing the original data.

The approach of sample-moment-based density approximant is an extension of the approach of moment-based density approximant, which is mathematically equivalent to traditional orthogonal polynomials approaches to the probability density function (Provost, 2005). This paper shows that, regardless of a negligible probability, a moment-based density approximant can be well approximated by its sample-moment-based approximant if the size of the sample used in the evaluation is reasonable large. Consequently, a density function can be reasonably approximated by its sample-moment-based density approximant.

This paper focuses on the properties and the performance of the approach of the sample-moment-based density approximant based on noise multiplied data. Due to the restriction on the number of pages, some technical issues on implementing the approach proposed in practice will be discussed in another paper.

Keywords: Confidential data, Masked data, Multiplicative noise, Moment-based density approximant.

1 Introduction

Many government institutions and statistical agencies collect survey data from individuals and businesses. Publishing these data with certain level of protection is necessary. Many different protection methods, including microaggregation of sensitive data, local suppression of unique data cells, top and bottom coding of continuous variables, rank swapping, rounding, adding noise, imputation and multiplicative noise, have been introduced and used in practice. More information on data protection can be found in Duncan and Lambert (1986 and 1989), Willenborg and De Waal (2001), Oganian (2010), Shlomo (2010), and the references therein.
The aim of government institutions and data agencies publishing the masked data sets is to provide end-users an opportunity to work out the statistical information on the underlying data without breaching confidentiality. As mentioned in Nayak et al. (2011), data perturbation may destroy unbiasedness and other properties of estimators. Methods and formulas for analysing an original data set may not be appropriate for analysing a masked version of it.

Describing and estimating the probability density function of a random variable are the basic tenets in statistical data analysis. Provost (2005) introduced the moment-based density approximant method for probability density approximation. He proved and demonstrated that using the moment-based density approximant to approach the density function is mathematically equivalent to using those orthogonal polynomials, such as the Legendre, Laguerre, Jacobi, and Hermite polynomials.

The multiplicative noise method is one type of noise addition used to perturb and protect confidential data. Kim and Jeong (2008) classified the multiplicative noise scheme into two schemes, Multiplicative Noise Scheme I and Multiplicative Noise Scheme II. The multiplicative noise scheme considered in this paper is Multiplicative Noise Scheme I. It is briefly defined as follows. Let \( Y \) be a sensitive random variable with observations \( y_1, y_2, \ldots, y_N \) (original data). Let \( C \) be a positive random variable, independent of \( Y \). When we say the original data \( y_1, y_2, \ldots, y_N \) are masked by \( C \), it means the masked data have the form \( y_i^* = y_i \times c_i \) where \( \{c_i\} \) is a sample from \( C \). In literature, sometimes it imposes \( E(C) = 1 \). With this restriction, \( y^* \) is an unbiased estimator of \( y \) given \( y \). This restriction does not apply to the method proposed in this paper. Therefore, the unbiased estimator of \( y \) will be \( y^*/E(C) \), given \( y \). Without further explanation, the term “masked data” used in this paper is for “noise multiplied data”.

For noise multiplied data, developing appropriate data analysis methods and formulas for different inference purposes is necessary (see Kim and Jeong (2008) for domain estimation, Sinha, et al. (2011) for quantile estimation, and Lin and Wise (2012) for linear regression parameters estimation). This paper proposes a method to obtain the density approximant of a sensitive random variable \( Y \) based on its masked data.

Many properties of the multiplicative noise method, including evaluation of disclosure risk, confidential protection, moment estimation, linear regression parameter estimation, properties of balanced noise distribution and effects on data quality and privacy protection in context of tabular magnitude data, have been deeply discussed and investigated in literature (Evans, 1996; Evans et al., 1998; Hwang, 1998; Kim and Winkler, 2003; Kim and Jeong, 2008; Oganian, 2010; Krsinich and Piesse, 2002; Nayak, et al., 2011; Sinha, et al., 2011; Lin and Wise, 2012 and Klein and Sinha, 2013). One of the important properties is the moments of \( Y \) can be evaluated through the moments of its masked variable \( Y^* \) and the moments of the noise \( C \) used to mask \( Y \).

With the well developed numerical result of the density approximant provided by Provost (2005) and the nice relationship among the moments of \( Y \), masked variable \( Y^* \) and noise \( C \), respectively, the density function of \( Y \) can be theoretically well approximated by the density approximant based on the moments