Unbiased density estimation for stochastically scaled Gaussian vectors using random Riemann sums

Patrick Muchmore **Paul Marjoram**
Department of Preventive Medicine
University of Southern California
Los Angeles, CA 90089
{muchmore,pmarjora}@usc.edu

Abstract

A range of probabilistic phenomena can be modeled as a Gaussian vector subject to a stochastic scaling by a positive random variable. While conceptually straightforward to define, the vast majority of cases result in random variables with density functions that are either unknown or computationally intractable. Here we consider the problem of density estimation for this family of random variables. We first derive a standard Monte Carlo estimator which reduces the problem to sampling from $\mathbb{R}^+$ and evaluating the multivariate Gaussian density. We then introduce an estimator based on random Riemann sums, which provides significantly higher accuracy for roughly the same computational cost. Because these estimators are unbiased, they can also form the basis of pseudo-marginal Markov chain Monte Carlo (MCMC) methods for this family, enabling exact inference despite an intractable, or even unknown, likelihood function.

1 Introduction

We start with a Gaussian vector $Y \sim N(\nu, \Sigma)$, where $\nu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$. For constant $\mu \in \mathbb{R}^p$ and univariate W, a positive $(P(W > 0) = 1)$ random variable independent of Y, we define the stochastically scaled Gaussian vector X to be

$$X := \mu + WY. \tag{1}$$

If $W \equiv 1$ then X is Gaussian, otherwise it is non-Gaussian. In the non-Guassian case if $\nu = 0$ the distribution of X is elliptically contoured, which is the multidimensional analogue of univariate symmetry [1–3]. If X is non-Gaussian and $\nu \neq 0$ the distribution is skew; the term *non-central* is also used in this case (and similarly *central* if $\nu = 0$). For identifiability purposes, when $W \equiv 1$ and X is Gaussian we always assume $\nu = 0$, which implies $X \sim N(\mu, \Sigma)$.

The Gaussian and (central) Student's t distributions are essentially the only two examples where the density of X can be expressed in closed form. To approximate the density f_X, in this paper we present three approaches to estimation: direct numerical integration, a standard Monte Carlo average, and a method based on the theory of random Riemann sums [4–6] (which can be seen as an example of stratified sampling [7–9]). Because we approach the question of calculating f_X as an integration problem, we will use the notation $\widehat{I}_{<\text{name}>}$ to denote the estimates produced by each method.

Ideally, we would like to be able to approximate f_X to arbitrary precision. If that is not feasible, exact inference is still possible via the "pseudo-marginal" approach to MCMC [10, 11] if the likelihood estimate is unbiased. Both the Monte Carlo and random Riemann sum estimators possess this quality, while numerical integration does not.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

1.1 Conditional, joint, and marginal densities

We assume Σ has full rank, so we can write the density function of Y as

$$f_Y\left(y|\,\nu,\Sigma\right) = (2\pi)^{-\frac{p}{2}}|L|^{-1}\exp(-1/2(y-\nu)^{\mathsf{T}}\Sigma^{-1}(y-\nu)) \tag{2}$$

where L is the unique lower triangular matrix such that $LL^{\mathsf{T}} = \Sigma$. Therefore, the conditional distribution $X|W = w \sim N(\mu + w\nu, w^2\Sigma)$ has density function

$$\begin{aligned} f_{X|W}(x|\nu,\Sigma,\mu,w) &= (2\pi)^{-\frac{p}{2}}|wL|^{-1}\exp\left(-1/2(x-\mu-w\nu)^{\mathsf{T}}(w^2\Sigma)^{-1}(x-\mu-w\nu)\right) \\ &= w^{-p} f_Y\left(\frac{x-\mu}{w}\middle|\,\nu,\Sigma\right). \end{aligned} \tag{3}$$

By definition, the joint density of X and W is the product $f_{X|W}f_W = f_{X,W}$, and the marginal density of X is the integral of this with respect to W

$$f_X(x|\nu,\Sigma,\mu) = \int_0^\infty f_{X|W}(x|\nu,\Sigma,\mu,w)f_W(w)dw = E_W(f_{X|W}). \tag{4}$$

2 Likelihood approximation methods

2.1 Numerical integration

One approach to estimating the likelihood (4) is to use numerical integration (NI) to calculate the integral of the product $f_{X|W}f_W = f_{X,W}$. This integral runs over the positive real numbers, and at least in principle it can be calculated numerically, for example in R [12] with the `integrate` function from the `stats` package. Therefore, the first estimator is defined to be

$$\widehat{I}_{\text{NI}} := \widehat{\int_0^\infty} f_{X|W}(x|\nu,\Sigma,\mu,w)f_W(w)dw, \tag{5}$$

where $\widehat{\int_a^b}$ denotes a (black-box) numerical integration routine. Equation (3) shows the conditional density $f_{X|W}$ can be calculated using the normal density (2), and an R implementation of the multivariate normal density is provided by the `mvtnorm` [13] package. However, this function recomputes the Cholesky decomposition of Σ every time it is called. Because $f_{X|W}$ is evaluated multiple times at a fixed value of Σ, it is more efficient to precompute L once and reuse it each time.

2.2 Monte Carlo averaging

A standard Monte Carlo (MC) estimate of (4) can be calculated by first simulating a realization w of the random variable W, and then calculating $f_{X|W}(x|\nu,\Sigma,\mu,w)$ at w as an approximation of the density $f_X(x|\nu,\Sigma,\mu)$. To improve the accuracy of the estimate, we can take the mean of n independent simulated realizations $w_1, w_2, \ldots w_n$, and the MC estimator is defined to be

$$\widehat{I}_{\text{MC}} := \frac{1}{n}\sum_{k=1}^{n} f_{X|W}(x|\nu,\Sigma,\mu,w_k). \tag{6}$$

Each individual MC value is an unbiased estimate of f_X, so by the linearity of expectation $\widehat{I}_{\text{MC}}$ is an unbiased estimator as well.

2.3 Random Riemann sums

The random Riemann sum estimator is similar to the MC approach, in that it involves generating n simulated random variables and evaluating $f_{X|W}$ at these values. However, instead of simulating independent replicates of W, the sample space is partitioned into n pieces of equal probability ($\frac{1}{n}$), and a single value is sampled from each. We will use the notation $w_{\frac{k}{n}}$ to indicate the realization from the kth partition, and we define the RRS estimator to be:

$$\widehat{I}_{\text{RRS}} := \frac{1}{n}\sum_{k=1}^{n} f_{X|W}(x|\nu,\Sigma,\mu,w_{\frac{k}{n}}). \tag{7}$$

In general, the problem of partitioning the sample space into n equiprobable pieces presents significant challenges; however, because W is restricted to $\mathbb{R}^+$ it is greatly simplified in this context. If the quantile function for W is known, or can be computed numerically, a partition consisting of n intervals I_k can be explicitly expressed as $I_k = \left[F_W^{-1}(\frac{k-1}{n}), F_W^{-1}(\frac{k}{n})\right)$, and the simulated values $w_{\frac{k}{n}}$ generated via inverse transform sampling. If the quantile function is unknown, techniques such as importance sampling can be employed to generate realizations from $W_{\frac{k}{n}}$.

The unbiasedness of (7) stems from the fact that $f_{W_{\frac{k}{n}}} = nf_W 1_{w\in I_k}$ and (4) can be written as $\sum_{k=1}^{n}\int_0^\infty f_{X|W} f_W 1_{w\in I_k} dw$, where $1_{w\in I_k}$ is the indicator function for the event $w \in I_k$. It can also be shown that $\mathrm{var}(\widehat{I}_{\mathrm{RRS}}) \leq \mathrm{var}(\widehat{I}_{\mathrm{MC}})$, with the reduction in variance a function of the difference between the partition means and the overall mean [9].

3 Student's t Example

The best known example when $W \not\equiv 1$ is the multivariate Student's t. This distribution arises from the model (1) by choosing $\nu = 0$ and $W \sim \sqrt{\frac{k}{\chi_k^2}}$, where χ_k^2 has a chi-squared distribution with k degrees of freedom. In this case the density function can be expressed in closed form as [14]

$$f_T(t|\Sigma,\mu,k) = \frac{\Gamma((k+p)/2)}{\Gamma(k/2)}(k\pi)^{-\frac{p}{2}}|L|^{-1}\left(1 + {}^1\!/\!_k(t-\mu)^\intercal\Sigma^{-1}(t-\mu)\right)^{-\frac{(k+p)}{2}}. \tag{8}$$

The relationship $W \sim \sqrt{\frac{k}{\chi_k^2}}$ enables us to generate a simulated realization of W by first generating a χ_k^2 random variate, then transforming the result. Similarly, we can use the definition of W to express its distribution, density, and quantile functions in terms of the corresponding functions for a χ_k^2 random variable. Specifically, the transformation $W = h(\chi_k^2) = \sqrt{\frac{k}{\chi_k^2}}$ is (strictly) monotone decreasing, with inverse $\chi_k^2 = h^{-1}(W) = \frac{k}{W^2}$. Therefore, the functions characterizing W can be expressed as: $f_W(w) = [dh^{-1}(w)/dw]f_{\chi_k^2}(h^{-1}(w)) = 2kw^{-3}f_{\chi_k^2}(kw^{-2})$, $F_W(w) = 1 - F_{\chi_k^2}(h^{-1}(w)) = 1 - F_{\chi_k^2}(kw^{-2})$, and $F_W^{-1}(p) = h(F_{\chi_k^2}^{-1}(1-p)) = \sqrt{k/F_{\chi_k^2}^{-1}(1-p)}$.

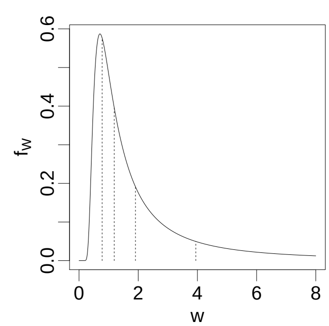

Figure 1: Quintiles of W for multivariate Student's t with $\mathrm{df} = 1$.

Because the multivariate Student's t density is known and computationally tractable, given (simulated) data x the (log) likelihood can be explicitly calculated. Therefore, we will use this distribution to compute exact likelihoods as a baseline for comparison with the results of each approximation. Implementations of the multivariate t density and a random number generator (RNG) are provided as part of the `mvtnorm` [13] R package.

Figure 2 illustrates the results of applying all three estimation methods to the problem of estimating a multivariate Student's t density with data vectors of size $p = \{10, 100, 1000\}$. In these examples μ is assumed to be a vector of 1s, while Σ is tridiagonal with 1 on the diagonal and $\frac{1}{2}$ on the sub/superdiagonal (ν must be a vector of 0s because this is an elliptically contoured distribution). Based on these parameter values, the data x consists of a simulated realization with density (8). Algorithms (6) and (7) also require specifying the number n of simulated W values to generate. In Figure 2 n runs from 50-2500, with density estimates calculated at increments of size 50.

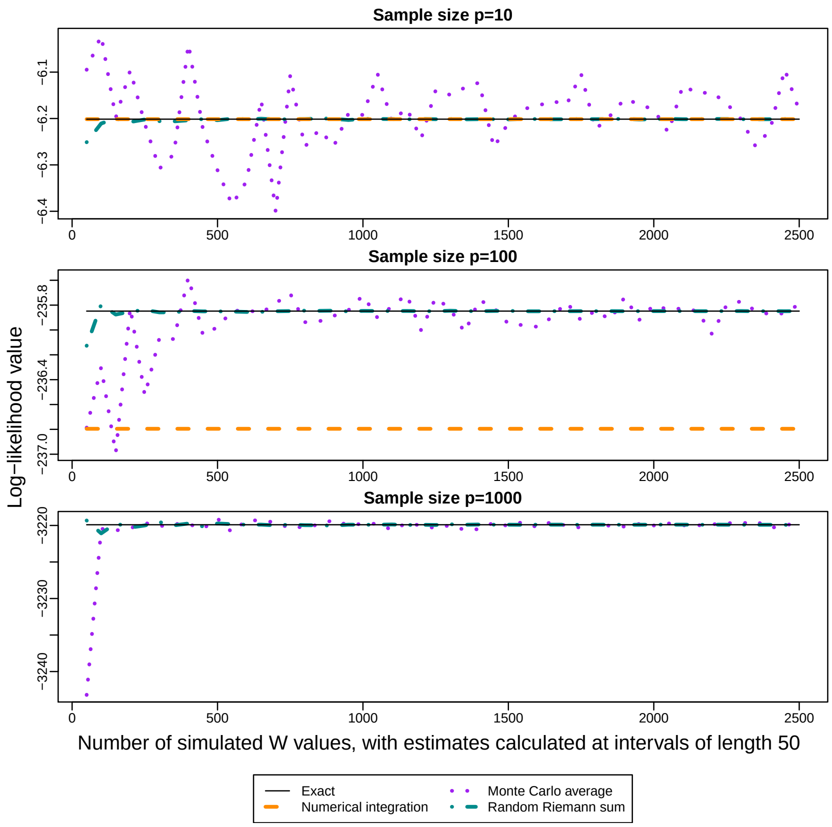

Figure 2: Three multivariate Student's t density estimators and exponentially increasing sample sizes.

With small sample sizes NI gives accurate results. Although, with a sample size of $p = 100$ there is significant error in the calculation, and with $p >\sim 100$ the calculation fails completely. As expected, both the MC and RRS estimators appear unbiased, with empirical evidence the RRS estimator has a significantly smaller variance.

4 Discussion

The class of elliptically contoured distributions, examples of which occur when $\nu = 0$, has been well characterized from a theoretical point of view; however, because most cases do not possess a closed form density statistical inference has proven difficult. An example of particular interest is the multivariate elliptical Laplace distribution, which has applications ranging from finance and engineering to biology and environmental science [15]. In this case the required scaling variable is $W = \sqrt{E}$, where E is standard exponential, a straightfoward variation of the Student's t example illustrated. Because this distribution has no closed form density MCMC sampling is not currently possible, but with the unbiased estimators developed here exact inference can still be accomplished in the pseudo-marginal framework.

When $\nu \neq 0$ the resulting family of distributions are skew (or non-central). In this case even Student's t has a density function with no closed form expression. For this reason these distributions have received much less attention; however, in many ways that increases their appeal as a target for these estimation methods. Because $\nu \neq 0$ corresponds to the alternative hypothesis in standard regression models, their most familiar application arises in the context of power calculations to determine sample sizes. Preliminary evidence indicates the estimators developed here provide significant improvements in accuracy over current methods of non-central density estimation. Moreover, because these estimators are unbiased, even in the non-central case they can be used for exact inference via pseudo-marginal MCMC.

References

[1] Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. Monographs on statistics and applied probability. Chapman and Hall, London; New York, 1990. ISBN 9780412314308.

[2] Kai-Tai Fang and Yao-Ting Zhang. *Generalized Multivariate Analysis*. Spring-Verlag, Berlin; New York, 1990. ISBN 9783540176510.

[3] Arjun K Gupta, Tamas Varga, and Taras Bodnar. *Elliptically Contoured Models in Statistics and Portfolio Theory*. Springer-Verlag, New York, 2nd edition, 2013. ISBN 9781461481539.

[4] John C Kieffer and Časlav V Stanojević. The Lebesgue integral as the almost sure limit of random riemann sums. *Proceedings of the American Mathematical Society*, 85(3):389–392, 1982.

[5] Alexander Pruss. Randomly sampled Riemann sums and complete convergence in the law of large numbers for a case without identical distribution. *Proceedings of the American Mathematical Society*, 124(3):919–929, 1996.

[6] Henar Urmeneta and Víctor Hernández. Random Riemann sum estimator versus Monte Carlo. *Computational Statistics & Data Analysis*, 51(9):4717–4730, 2007.

[7] Keith D Tocher. *The Art of Simulation*. English Universities Press, London, 1963.

[8] William G Cochran. *Sampling Techniques*. Wiley, New York, 3rd edition, 1977. ISBN 047116240X.

[9] Michael D McKay, Richard J Beckman, and William J Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

[10] Mark A Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

[11] Christophe Andrieu and Gareth O Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[13] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2017. URL `https://CRAN.R-project.org/package=mvtnorm`. R package version 1.0-6.

[14] Samuel Kotz and Saralees Nadarajah. *Multivariate T-Distributions and Their Applications*. Cambridge University Press, Cambridge; New York, 2004. ISBN 9780521826549.

[15] Samuel Kotz, Tomasz Kozubowski, and Krzystof Podgorski. *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering, and Finance*. Progress in Mathematics Series. Springer Science & Business Media, New York, 2001. ISBN 9780817641665.