Applicable Analysis and Discrete Mathematics
available online at http://pefmath.etf.rs

Appl. Anal. Discrete Math. **13** (2019), 871–882.
https://doi.org/10.2298/AADM181222039S

COMPUTING TAIL AREAS FOR A HIGH-DIMENSIONAL GAUSSIAN MIXTURE

Dedicated to Academician Professor Gradimir Milovanović on the occasion of his 70th birthday.

*Burcin Simsek, Satish Iyengar**

We consider the problem of computing tail probabilities — that is, probabilities of regions with low density — for high-dimensional Gaussian mixtures. We consider three approaches: the first is a bound based on the central and non-central χ^2 distributions; the second uses Pearson curves with the first three moments of the criterion random variable U; the third embeds the distribution of U in an exponential family, and uses exponential tilting, which in turn suggests an importance sampling distribution. We illustrate each method with examples and assess their relative merits.

1. INTRODUCTION

Suppose that X has the following finite Gaussian mixture probability density function (pdf) in $\mathbb{R}^d$:

$$f(x) = \sum_{i=1}^{c} \gamma_i \phi(x|\mu_i, \Sigma_i), \tag{1}$$

where

$$\phi(x|\mu, \Sigma) = \phi_d(x|\mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right],$$

*Corresponding author. Satish Iyengar
2010 Mathematics Subject Classification. Primary 60E05; 62E20; 41A60; 33C20.
Keywords and Phrases. Chi-square, exponential family, exponential tilting, non-central chi-square, Pearson curves.

is the Gaussian pdf with mean μ and covariance matrix Σ, $|2\pi\Sigma|$ is the determinant of $2\pi\Sigma$, and the mixing weights are $\gamma_i \geq 0$ with $\sum_i \gamma_i = 1$. Having observed $[X = a]$, we address the problem of estimating the tail probability

$$p_t = P[U = f(X) \leq f(a) = t]. \tag{2}$$

We call U the criterion (random) variable, and denote its pdf and cdf as $g(u)$ and $G(u)$, respectively. Such problems arise in several contexts: for example, see [**3, 7**] for genetic and psychiatric applications of mixture distributions. More generally, as the sizes of data sets increase, it is often the case that the data are a mixture of several homogeneous distributions rather than a single one. And in a streaming data environment, the occurrence of an observation that is highly unlikely under any of the current component pdfs $f_i(x) = \phi(x|\mu_i, \Sigma_i)$ may signal the emergence of a new cluster. Thus, given an observation from a mixture distribution, deciding whether a particular observation is unusual or not requires the evaluation of (2).

In this paper, we suggest three approaches for this problem: in Section , we derive bounds based on the central and non-central χ^2 distributions; in Section we show how to compute and use the moments of $U = f(X)$ to estimate it; in Section , we use exponential tilting instead; in Section we illustrate these approaches with examples and end with a discussion of the methods' relative merits.

2. χ^2 BOUNDS

Let M be the mixing random variable with $P(M = i) = \gamma_i$, and write $f_i(x) = \phi(x|\mu_i, \Sigma_i)$ for the pdf of the i^{th} component. The probability of interest is

$$\begin{aligned} p_t &= \sum_{i=1}^{c} \gamma_i P[f(X) \leq t | M = i] = \sum_{i=1}^{c} \gamma_i P\left[\sum_{j=1}^{c} \gamma_j f_j(X) \leq t | M = i\right] \\ &\leq \sum_{i=1}^{c} \gamma_i P\left[\gamma_i f_i(X) \leq t | M = i\right] = \sum_{i=1}^{c} \gamma_i P\left[f_i(X) \leq \gamma_i^{-1} t | M = i\right]. \end{aligned}$$

Now let $Q_i = (X - \mu_i)'\Sigma_i^{-1}(X - \mu_i)$ and let $t > 0$. Then

$$P[f_i(X) \leq u | M = i] = P\left(Q_i \geq -2 \ln\left[u|2\pi\Sigma_i|^{1/2}\right] | M = i\right).$$

Given $[M = i]$, the conditional distribution of Q_i is χ^2_d; thus, the upper bound in

$$p_t \leq \sum_{i=1}^{c} \gamma_i P\left(Q_i \geq -2 \ln \frac{t|2\pi\Sigma_i|^{1/2}}{\gamma_i}\right), \tag{3}$$

is easy to compute. The performance of (3) will be assessed in Section .

The inequality above is valid because a sum of nonnegative numbers is at most t implies that each summand is at most t. The choice of $j = i$ gives a chi-square expression; if we choose $j \neq i$, we have a more involved calculation. In particular, we have

$$\begin{aligned} P(f(X) \leq t) &= \sum_{i=1}^{c} \gamma_i P[f(X) \leq t | M = i] = \sum_{i=1}^{c} \gamma_i P\left[\frac{1}{c}\sum_{j=1}^{c} \gamma_j f_j(X) \leq \frac{t}{c} | M = i\right] \\ &\leq \sum_{i=1}^{c} \gamma_i \sum_{j=1}^{c} P\left[\gamma_j f_j(X) \leq \frac{t}{c} | M = i\right] \qquad (4) \\ &= \sum_{i=1}^{c} \gamma_i \sum_{j=1}^{c} P\left[Q_{ij} \geq -2\ln \frac{t|2\pi\Sigma_j|^{1/2}}{\gamma_j c} | M = i\right], \end{aligned}$$

where $Q_{ij} = (X - \mu_j)'\Sigma_j^{-1}(X - \mu_j)$ when $[M = i]$ and X is a $N(\mu_i, \Sigma_i)$ random vector. This because if the average of positive numbers is less than t/c, then at least one of them must be less than t/c. Now, conditional on $[M = i]$, X is a $N(\mu_i, \Sigma_i)$ random vector which we can represent as $X = \Sigma_i^{1/2} Z + \mu_i$, where Z is a $N(0, I)$ random vector. Thus,

$$\begin{aligned} Q_{ij} &= (\Sigma_i^{1/2} Z + \mu_i - \mu_j)'\Sigma_j^{-1}(\Sigma_i^{1/2} Z + \mu_i - \mu_j) \\ &= \left(Z + \Sigma_i^{-1/2}(\mu_i - \mu_j)\right)' \Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2} \left(Z + \Sigma_i^{-1/2}(\mu_i - \mu_j)\right). \qquad (5) \end{aligned}$$

Using the spectral decomposition $\Sigma_i^{1/2}\Sigma_j^{-1}\Sigma_i^{1/2} = \Gamma_{ij}' D_{ij} \Gamma_{ij}$, conditional on $[M = i]$ we have

$$Q_{ij} \stackrel{d}{=} (Z + \nu_{ij})' D_{ij} (Z + \nu_{ij}),$$

where $\nu_{ij} = \Gamma_{ij}\Sigma_i^{-1/2}(\mu_i - \mu_j)$, and $\stackrel{d}{=}$ means having the same distribution. Finally, just as in (4) we have another upper bound

$$P(U \leq t) \leq \sum_{i=1}^{c}\sum_{j=1}^{c} \gamma_i P\left(Q_{ij} \geq -2\ln \frac{t|2\pi\Sigma_j|^{1/2}}{\gamma_j c}\right). \qquad (6)$$

The distribution of Q_{ij} is a weighted sum of noncentral χ_1^2 random variables, with weights equal to the eigenvalues of D_{ij}. This bound in (6) involves a double sum, but each term is small because t is replaced by t/c and it involves the probabilities of a $N(\mu_i, \Sigma_i)$ random vector being outside an ellipsoidal contour of a $N(\mu_j, \Sigma_j)$ distribution, which can be quite small for well separated components. However, because the distribution of Q_{ij} is difficult, we will not assess this bound in Section 5.

Due to the poor performance of both (3) and (6) bounds, we argue for other methods that can provide better approximations, to which we turn next.

3. PEARSON CURVE APPROXIMATION

The cumulative distribution function (cdf) of U appears to be intractable; however, we will see below that all of its moments are tractable because they can be expressed in terms of the Gaussin pdf at 0. Because U is a bounded (nonnegative) random variable, its moments determine its distribution. Now, if the parameters $\{(\gamma_i, \mu_i, \Sigma_i) : i = 1, \ldots, c\}$ are all known, classical methods due to Chebyshev and others can bound the cdf of U using those moments [**8**]. Such bounds can be hard to compute when using many moments; and even though these bounds are best in the tails, they can be quite loose, at least in relative terms [**6**]. On the other hand, if the parameters are estimated from data, then it is prudent to use only low order moments of U because the higher order ones are typically highly variable. In particular, we can use Pearson curves [**9**] to approximate the distribution of U: in general, four moments are needed to do a Pearson fit; however, for a nonnegative random variable, only the first three moments along with the lower endpoint at the origin suffice. We will see below that the number of terms needed to compute the first three moments is on the order of c^4, so they are tractable unless c is very large.

For the problem at hand, let let $\hat{G}_3(u)$ be the Pearson curve approximation to $G(u)$ based on the first three moments. Of course, comparing the cdfs is equivalent to comparing the quantiles $G^{-1}(p)$ with those of $\hat{G}_3^{-1}(p)$. We next compute the first three moments of U.

The moments of $U = f(X)$ are

$$E[U^n] = E[f(X)^n] = \int_{\mathbb{R}^d} f(x)^{n+1}\, dx. \tag{7}$$

The integrand in (7) is a polynomial in Gaussian pdfs. Fortunately, products and powers of those pdfs are proportional to other Gaussian pdfs. In particular, upon completing the square we have

$$\prod_{i=1}^{p} \phi(x|\nu_i, C_i) = \frac{\prod_{i=1}^{p} \phi(0|\nu_i, C_i)}{\phi(0|\nu, C)} \phi(x|\nu, C), \tag{8}$$

where

$$C^{-1} = \sum_{i=1}^{p} C_i^{-1} \quad \text{and} \quad \nu = C \sum_{i=1}^{p} C_i^{-1} \nu_i.$$

Such expressions also appear in product of experts models [**5**], for example in speech recognition [**1**]. Integrating such products is easy:

$$\begin{aligned} \int_{\mathbb{R}^d} \prod_{i=1}^{p} \phi(x|\nu_i, C_i)\, dx &= \frac{\prod_{i=1}^{p} \phi(0|\nu_i, C_i)}{\phi(0|\nu, C)} \int_{\mathbb{R}^d} \phi(x|\nu, C)\, dx \\ &= \frac{\prod_{i=1}^{p} \phi(0|\nu_i, C_i)}{\phi(0|\nu, C)}. \end{aligned} \tag{9}$$

When $\nu_i \equiv \nu_0$ and $C_i \equiv C_0$,

$$\phi(x|\nu_0, C_0)^p = \frac{\phi(0|\nu_0, C_0)^p}{\phi(0|\nu_0, C_0/p)}\phi(x|\nu_0, C_0/p) = \frac{\phi(0|0, C_0)^p}{\phi(0|0, C_0/p)}\phi(x|\nu_0, C_0/p); \tag{10}$$

note that (10) is valid for all $p > 0$, not just integers. Using these formulas, we now present explicit expressions for the first three moments of U.

First moment. For the parameters in (1) and (2), let $\alpha_i = \gamma_i\phi(x|\mu_i, \Sigma_i)$. Using the identity

$$\left(\sum_{i=1}^{c}\alpha_i\right)^2 = \sum_{i=1}^{c}\alpha_i^2 + 2\sum_{i<j}\alpha_i\alpha_j$$

and equations (4), (5), and (6), the mean of U is

$$E(U) = \sum_{i=1}^{c}\gamma_i^2\frac{\phi(0|0, \Sigma_i)^2}{\phi(0|0, \Sigma_i/2)} + 2\sum_{i<j}\gamma_i\gamma_j\frac{\phi(0|\mu_i, \Sigma_i)\phi(0|\mu_j, \Sigma_j)}{\phi(0|\mu_{ij}, \Sigma_{ij})}, \tag{11}$$

where

$$\Sigma_{ij}^{-1} = \Sigma_i^{-1} + \Sigma_j^{-1} \quad \text{and} \quad \mu_{ij} = \Sigma_{ij}(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j).$$

Second moment. Next, using the identity

$$\left(\sum_{i=1}^{c}\alpha_i\right)^3 = \sum_{i=1}^{c}\alpha_i^3 + 3\sum_{i\neq j}\alpha_i^2\alpha_j + 6\sum_{i<j<k}\alpha_i\alpha_j\alpha_k$$

the second moment of U is

$$\begin{aligned} E(U^2) &= \sum_{i=1}^{c}\gamma_i^3\frac{\phi(0|0, \Sigma_i)^3}{\phi(0|0, \Sigma_i/3)} + 3\sum_{i\neq j}\gamma_i^2\gamma_j\frac{\phi(0|\mu_i, \Sigma_i)^2\phi(0|\mu_j, \Sigma_j)}{\phi(0|\mu_{ij}^{21}, \Sigma_{ij}^{21})} \\ &\quad + 6\sum_{i<j<k}\gamma_i\gamma_j\gamma_k\frac{\phi(0|\mu_i, \Sigma_i)\phi(0|\mu_j, \Sigma_j)\phi(0|\mu_k, \Sigma_k)}{\phi(0|\mu_{ijk}^{111}, \Sigma_{ijk}^{111})} \end{aligned} \tag{12}$$

where

$$[\Sigma_{ij}^{21}]^{-1} = (2\Sigma_i^{-1} + \Sigma_j^{-1}), \quad \mu_{ij}^{21} = \Sigma_{ij}^{21}(2\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j),$$

$$[\Sigma_{ijk}^{111}]^{-1} = (\Sigma_i^{-1} + \Sigma_j^{-1} + \Sigma_k^{-1}), \quad \text{and} \quad \mu_{ijk}^{111} = \Sigma_{ijk}^{111}(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k).$$

Third moment. Finally, using the identity

$$\left(\sum_{i=1}^{c}\alpha_i\right)^4 = \sum_{i=1}^{c}\alpha_i^4 + 4\sum_{i\neq j}\alpha_i^3\alpha_j + 6\sum_{i<j}\alpha_i^2\alpha_j^2 + 12\sum_{j<k,i\neq j,k}\alpha_i^2\alpha_j\alpha_k + 24\sum_{i<j<k<l}\alpha_i\alpha_j\alpha_k\alpha_l,$$

the third moment of U is

$$\begin{aligned} E(U^3) &= \sum_{i=1}^{c}\gamma_i^4\frac{\phi(0|0,\Sigma_i)^4}{\phi(0|0,\Sigma_i/4)} + 4\sum_{i\neq j}\gamma_i^3\gamma_j\frac{\phi(0|\mu_i,\Sigma_i)^3\phi(0|\mu_j,\Sigma_j)}{\phi(0|\mu_{ij}^{31},\Sigma_{ij}^{31})} \\ &\quad + 6\sum_{i<j}\gamma_i^2\gamma_j^2\frac{\phi(0|\mu_i,\Sigma_i)^2\phi(0|\mu_j,\Sigma_j)^2}{\phi(0|\mu_{ij}^{22},\Sigma_{ij}^{22})} \\ &\quad + 12\sum_{i\notin\{j,k\};j<k}\gamma_i^2\gamma_j\gamma_k\frac{\phi(0|\mu_i,\Sigma_i)^2\phi(0|\mu_j,\Sigma_j)\phi(0|\mu_k,\Sigma_k)}{\phi(0|\mu_{ij}^{211},\Sigma_{ij}^{211})} \\ &\quad + 24\sum_{i<j<k<l}\gamma_i\gamma_j\gamma_k\gamma_l\frac{\phi(0|\mu_i,\Sigma_i)\phi(0|\mu_j,\Sigma_j)\phi(0|\mu_k,\Sigma_k)\phi(0|\mu_l,\Sigma_l)}{\phi(0|\mu_{ijkl}^{1111},\Sigma_{ijkl}^{1111})}, \end{aligned} \tag{13}$$

where

$$[\Sigma_{ij}^{31}]^{-1} = (3\Sigma_i^{-1} + \Sigma_j^{-1}), \quad \mu_{ij}^{31} = \Sigma_{ij}^{31}(3\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j),$$

$$[\Sigma_{ij}^{22}]^{-1} = (2\Sigma_i^{-1} + 2\Sigma_j^{-1}), \quad \mu_{ij}^{22} = \Sigma_{ij}^{22}(2\Sigma_i^{-1}\mu_i + 2\Sigma_j^{-1}\mu_j),$$

$$[\Sigma_{ijk}^{211}]^{-1} = (2\Sigma_i^{-1} + \Sigma_j^{-1} + \Sigma_k^{-1}), \quad \mu_{ijk}^{211} = \Sigma_{ijk}^{211}(2\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k),$$

$$[\Sigma_{ijkl}^{1111}]^{-1} = (\Sigma_i^{-1} + \Sigma_j^{-1} + \Sigma_k^{-1} + \Sigma_l^{-1}), \text{ and}$$

$$\mu_{ijkl}^{1111} = \Sigma_{ijkl}^{1111}(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j + \Sigma_k^{-1}\mu_k + \Sigma_l^{-1}\mu_l).$$

Using these moments, we use the program provided by Davis and Stephens [**4**] to compute the quantiles of the corresponding Pearson curve: see Section 5.

4. EXPONENTIAL TILTING AND IMPORTANCE SAMPLING

In this section, we embed the distribution of U in an exponential family and then use exponential tilting to suggest an importance sampling procedure [**2**]. We first show that the resulting procedure has smaller variance than simple Monte Carlo; we then deal with the problems of computing the appropriate importance sampling distribution and how to generate samples from it. These computational details can be cumbersome; however, if in practice the parameters of the mixture are not changing, then these computations should be tractable for routine use.

Let $K(\theta) = \ln E(e^{\theta U})$ be the cumulant generating function (cgf) of U, and construct the exponential family $\{g_\theta(u) : \theta \in \mathbb{R}\}$, where

$$g_\theta(u) = e^{\theta u - K(\theta)} g(u) \tag{14}$$

is the pdf of the tilted distribution; note that $g_0 = g$. Let P_θ and E_θ indicate the probability and expectation corresponding to g_θ. We need the following facts. Since U is a bounded positive random variable, $K(\theta)$ exists for all θ. As $\theta \to -\infty$, the tilted distribution concentrates near zero. Note that $K'(\theta) = E_\theta(U)$ and $K''(\theta) = \mathrm{var}_\theta(U)$ are both positive, so $K(\theta)$ is strictly increasing and convex. We assume that $f(a) = t < E_0(U)$, because we are interested in low-probability events; thus, the solution to $K'(\theta) = t$ is $\theta_t < 0$.

Now let V have pdf g_{θ_t}, so that $E_{\theta_t}(V) = t$ by construction. From the expressions

$$p_t = P(U \le t) = E_0[I(U \le t)]$$

and

$$p_t = P(U \le t) = \int_0^t e^{-\theta_t v + K(\theta_t)} g_{\theta_t}(v)\,dv = E_{\theta_t}\left[e^{-\theta_t V + K(\theta_t)} I(V \le t)\right]$$

we have the following two Monte Carlo estimates of p_t:

$$\hat{p}_t = \frac{1}{B}\sum_{b=1}^{B} I(U_b \le t) \quad \text{and} \quad \hat{q}_t = \frac{1}{B}\sum_{b=1}^{B} e^{-\theta_t V_b + K(\theta_t)} I(V_b \le t). \tag{15}$$

Both estimators are unbiased, so to compare them we consider their second moments with a simulation sample size of $B = 1$:

$$E_0(\hat{p}_t^2) = p_t.$$

and

$$E_{\theta_t}(\hat{q}_t^2) = \int_0^t e^{-2\theta_t v + 2K(\theta_t)} g_{\theta_t}(v)\,dv = \int_0^t e^{-\theta_t v + K(\theta_t)} g(v)\,dv. \tag{16}$$

Now consider the exponent $[K(\theta_t) - \theta_t v]$ here. For $0 < v < t$, we have

$$K(\theta_t) < K(\theta_t) - \theta_t v < K(\theta_t) - \theta_t t = K(\theta_t) - \theta_t K'(\theta_t) \tag{17}$$

because $\theta_t < 0$. The upper bound in (17) has the form

$$h(\theta) = K(\theta) - \theta K'(\theta),$$

for which $h'(\theta) = -\theta K''(\theta)$. Since $K''(\theta)$ is positive, h has a maximum of $h(0) = 0$ at $\theta = 0$ and is strictly decreasing as θ moves away in either direction. In short, $h(\theta) < 0$ for $\theta \neq 0$, so from (16) we have

$$E(\hat{q}_t^2) = \int_0^t e^{-\theta_t v + K(\theta_t)} g(v)\,dv \le e^{h(\theta_t)} \int_0^t g(v)\,dv < p_t. \tag{18}$$

Thus, the second moment (and hence the variance and coefficient of variation) of $\hat{p}_t$ is smaller than that of $\hat{q}_t$.

We end this section with a few comments about the exponential tilting method for approximating tail probabilities.
1. It is easy to simulate from the tilted distribution when $\theta < 0$. First generate X from the Gaussian mixture in (1); next, construct the criterion variable $U = f(X)$; finally, retain it with probability $e^{\theta U}$. The resulting random variable will have pdf g_θ. Note, however, that when θ is much less than zero, the retention probability is typically quite small, making this scheme potentially inefficient. Thus, although $\hat{q}_t$ has smaller variance than $\hat{q}_t$, the inefficient sampling scheme could well make the simpler $\hat{p}_t$ competitive.

2. To determine θ_t, we must solve the equation

$$t = K'(\theta) = E_\theta(U) = \frac{E(Ue^{\theta U})}{E(e^{\theta U})}.$$

If the parameter values of the mixture do not vary over a time period of interest, we can compute and store $K'(\theta)$ for various values of $\theta < 0$ using Monte Carlo; then for an observed value of a we could use those stored values to bound the search and then solve for θ_t using elementary algorithms such as bisection or golden section search.

3. Next, although exponential tilting suggests $\hat{q}_t$, additional information about the left tail of g may suggest even better importance sampling procedures. Consider the following analog: the right tail of the standard normal Z, $\tau_a = P(Z > a) = \Phi(-a)$ for large a. The simple Monte Carlo estimate is

$$\hat{\tau}_1 = I(Z > a).$$

A useful function in this context is Mills' ratio $M(a) = P(Z \geq a)/\phi(a) = \Phi(-a)/\phi(a)$, which is a decreasing function of a; and for $a > 0$ it satisfies

$$\frac{a}{1+a^2} < M(a) < \frac{1}{a}, \quad \text{so that} \quad \tau_a \sim \frac{\phi(a)}{a}. \tag{19}$$

Using these inequalities, we can show that the squared coefficient of variation of this estimate is $CV^2(\hat{\tau}_1) = (1-\tau_a)/\tau_a \sim \sqrt{2\pi}ae^{a^2/2}$ as $a \to \infty$; thus the relative error grows rapidly with a. Next, we move the distribution to have mean a by an exponential tilt: the expression

$$\tau_a = \int_a^\infty \frac{\phi(x)}{\phi(x-a)}\phi(x-a)dx = \int_0^\infty e^{-ax-a^2/2}\phi(x)dx$$

yields the estimator $\hat{\tau}_2 = e^{-aZ-a^2/2}I(Z \geq 0)$ for which $E(\hat{\tau}_2^2) = e^{a^2}\Phi(-2a)$. Thus, $CV^2(\hat{\tau}_2) \sim a\sqrt{\pi/2}$ as $a \to \infty$, so that the relative error grows much more slowly

than for simple Monte Carlo.

However, consider the following heuristic argument. If X has pdf f, l'Hôpital's rule says that with suitable regularity the asymptotic behavior of $P(X > a)/f(a)$ is the same as that of $r(\alpha) = -f(a)/f'(a)$. Assuming such regularity, the integral on the right in

$$\int_a^\infty f(x)dx = r(a)f(a)\int_0^\infty \frac{f(x+a)}{r(a)f(a)}dx \tag{20}$$

approaches 1 as $a \to \infty$. Because it is bounded away from 0 it can be estimated readily with good relative accuracy. For the normal distribution $r(a) = 1/a$, and (20) is

$$\int_a^\infty \phi(x)dx = \frac{\phi(a)}{a}\int_0^\infty \frac{a\phi(x+a)}{\phi(a)}dx = \frac{\phi(a)}{a}\int_0^\infty e^{-x^2/2}ae^{-ax}dx;$$

hence the estimator $\hat{\tau}_3 = a^{-1}\phi(a)e^{-T^2/2}$ where T has an exponential pdf with mean $1/a$. In this case, the Mill's ratio inequalities show that $CV^2(\hat{\tau}_3) \sim 2/a^2$ as $a \to \infty$. Thus, $CV^2(\hat{\tau}_3)$ tends to zero as a increases so that the estimator's *relative* accuracy improves as the probability tends to zero. It is easy to see that this estimator also results from the (importance) sampling pdf $h(t) = ae^{-a(t-a)}$ for $t \geq a$.

Returning to the importance sampling for the Gaussian mixture problem, as calculations are done for various values of t, the information we gain about the left tail of U may well point to an importance sampling pdf (such as the exponential for the Gaussian tail) that is much easier to generate from, making the resulting estimate more efficient and more accurate than $\hat{\tau}_1$.

5. AN EXAMPLE

We now assess the accuracy of the χ^2 bound, the Pearson curve approximation, and the exponential tilting method with one example; we do not assess the bound based on the non-central χ^2 distribution because it is rather unweildy. The parameter values are the following: first, the dimension, number of components and weights are

$$d = 3, \quad c = 7, \quad \gamma_i = \frac{8-i}{28}, \text{ respectively.}$$

The seven component means are

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
0	1	-1	0	0	0	0
0	0	0	1	-1	0	0
0	0	0	0	0	1	-1

and the covariance matrices are

$$\Sigma_1 = \Sigma_4 = 0.2I, \quad \Sigma_2 = \Sigma_5 = 0.2I + 0.05J,$$

$$\Sigma_3 = \Sigma_6 = 0.2I - 0.05J, \quad \Sigma_7 = 0.2I + 0.1J.$$

where I is the identity matrix and $J = [J_{ij}]$ with $J_{ij} \equiv 1$.

χ^2 bound in equation (3):

For the tail probabilities of 0.01, 0.025, 0.05, and 0.1, we obtain the values of the χ^2 bound in equation and summarize them below:

χ^2 UPPER BOUND FOR THE TAIL PROBABILITIES

t	0.002	0.007	0.014	0.028
Upper bound of p_t in (3)	0.04622	0.13789	0.24474	0.41277

Note that *pchisq* with *lower.tail=FALSE* command in R package is used to obtain χ^2 upper bound for the given tail probabilities. As stated in Section 2, χ^2 bound in equation (3) performs poorly, so we propose the following approximations.

Pearson curve approximation:

For the probabilities given in the table below, we used the program in [**4**] to compute the quantiles using Pearson curves. We then used simulation with 100,000 replicates to estimate the quantiles directly.

QUANTILE ESTIMATES FROM PEARSON CURVES (PC) AND SIMULATION

probability	0.01	0.025	0.05	0.1
PC quantile	0.00140	0.00313	0.00665	0.01321
Simulated quantile	0.00101	0.00274	0.00559	0.01209

The Pearson curve approximations appear to fairly close to the simulation estimates; of course, more work is needed to assess the Pearson curve estimate for smaller probabilities.

Exponential tilting

We will first outline the algorithm to generate random variables from an exponentially tilted distribution given one from the normal mixture for $\theta > 0$:

- Draw a random variable from Gaussian mixture probability density function f(x) described in equation (1) assuming $N = 10000$. From these random sample, compute the bounded positive random variables u_i where $U_i = f(X_i)$.
- Obtain quantiles of U correspondence to 0.01, 0.025, 0.05, and 0.1. Store them q01, q025, q05, and q10. Now, let $t1 = q01$, $t2 = q025$, $t3 = q05$, and $t4 = q10$.
- Next, for $i = 1, 2, 3, 4$ find values of θ_i so that $t_i = K'(\theta_i) = \frac{E(Ue^{\theta_i U})}{E(e^{\theta_i U})}$.

- Now, flip a coin with probability $e^{\theta u_i}$ for $u_i > 0$ and $\theta < 0$. If the coin comes up heads, keep the random variable; otherwise toss it out.
- Use the random variables that are retained to calculate the Monte Carlo (MC) quantile estimate of p_t: $\hat{q}_t = \frac{1}{B}\sum_{b=1}^{B} e^{-\theta_t V_b + K(\theta_t)} I(V_b \le t)$.

Quantile estimates from Exponential Tilting

probability	0.01	0.025	0.05	0.1
MC quantile estimate	0.00164	0.00395	0.00689	0.01388
Simulated quantile	0.00101	0.00274	0.00559	0.01209

6. CONCLUSION

Three approaches have been used to address the problem of computing tail probabilities of Gaussian mixtures. The numerical example showed that approximation based on χ^2 bound performs quite poorly, as expected; Pearson curve approximation and exponential tilting approach provides better approximation.

REFERENCES

1. S.S. Airey, M.J.F. Gales: *Product of Gaussians as a distributed representation for speech recognition.* In EUROSPEECH-2003 (2003), 877 – 880.
2. O.E. Barndorff-Nielsen, D.R. Cox: *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, 1989.
3. J. Chen, P. Li: *Hypothesis test for normal mixture models: the EM approach.* Annals of Statistics, **37** (2009), 2523–2542.
4. C.S. Davis, M.A. Stephens: *Approximate percentage points using Pearson curves.* Applied Statistics, **32** (1983), 322 – 327.
5. G.E. Hinton: *Products of experts.* Proceedings of the Ninth International Conference on Artificial Neural Networks, **1** (1999), 1 – 6.
6. B.G. Lindsay, P. Basak: *Moments determine the tail of a distribution (but not much else).* American Statistician, **54** (2000), 248 – 251.
7. S.E. Medland, J.E. Schmitt, B.T. Webb, P.H. Kuo, M.C. Neale: *Efficient calculation of empirical P-values for genome-wide linkage analysis through weighted permutation.* Behavioral Genetics, **39** (2009), 91 – 100.
8. H.J. Royden : *Bounds on a distribution function when its first n moments are given.* Annals of Mathematical Statistics, **3** (1953), 361 – 376.
9. H. Solomon, M.A. Stephens : *Approximations to density functions using Pearson curves.* J. American Statistical Association, **73** (1978), 153 – 160.

Burcin Simsek
University of Pittsburgh
Department of Statistics
230 S Bouquet St.
Pittsburgh, PA 15213
E-mail: *bus5@pitt.edu*

(Received 22.12.2018)
(Revised 02.04.2019)

Satish Iyengar
University of Pittsburgh
Department of Statistics
Proffessor and Interim Chair
230 S Bouquet St.
Pittsburgh, PA 15213
E-mail: *ssi@pitt.edu*