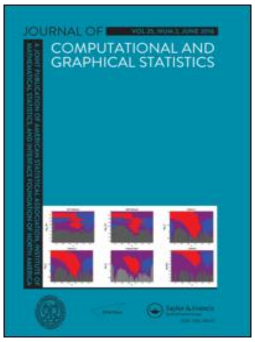

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Discrete Approximation of a Mixture Distribution via Restricted Divergence

Christian Röver & Tim Friede

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 1, 217–222
http://dx.doi.org/10.1080/10618600.2016.1276840

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Discrete Approximation of a Mixture Distribution via Restricted Divergence

Christian Röver and Tim Friede

Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

ABSTRACT
Mixture distributions arise in many application areas, for example, as marginal distributions or convolutions of distributions. We present a method of constructing an easily tractable discrete mixture distribution as an approximation to a mixture distribution with a large to infinite number, discrete or continuous, of components. The proposed DIRECT (divergence restricting conditional tesselation) algorithm is set up such that a prespecified precision, defined in terms of Kullback–Leibler divergence between true distribution and approximation, is guaranteed. Application of the algorithm is demonstrated in two examples. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2016
Revised October 2016

KEYWORDS
Convolution; DIRECT; Discrete approximation; Mixture distribution

1. Introduction

Mixture distributions with a large to infinite number of mixture components commonly occur in many fields of application (e.g., Seidel 2010). Common examples include, for example, marginal (posterior) distributions, convolutions of random variables, predictive distributions, distributions of test statistics, overdispersed sampling distributions, and many more.

If the mixture distribution's exact marginal density, distribution, or quantile functions are not available in analytical form, then practical application of such mixtures is often very limited. Such mixtures may then often be approximated to a sufficient degree by a mixture of a lower, *finite* number of components. How exactly to select such a finite set of components however is not obvious. In the following we describe a general approach and an algorithm allowing to set up a finite mixture as an approximation to a mixture distribution with a large or infinite number of components in a completely automated way. The construction is based on the *Kullback–Leibler divergence* or *relative entropy* between distributions and as such aims at bounding the (expected) logarithmic ratio of exact and approximate probability densities.

2. Kullback–Leibler Divergence

2.1. Definitions

The *Kullback–Leibler divergence* or *relative entropy* of two probability distributions with probability density functions p and q is defined as the *expected logarithmic ratio of densities* with respect to the former distribution (p),

$$\mathcal{D}_{\text{KL}}(p(\theta)\|q(\theta)) = \int_{\Theta} \log\left(\frac{p(\theta)}{q(\theta)}\right) p(\theta)d\theta = \text{E}_{p(\theta)}\left[\log\left(\frac{p(\theta)}{q(\theta)}\right)\right] \quad (1)$$

(Cover and Thomas 1991, chap. 2). In case of discrete probability distributions p and q, the integrals simplify to sums, but for simplicity we will stick to the integral notation in the following. The relative entropy is always positive, it is zero if the two distributions are identical ($p = q$), and larger otherwise. The divergence (in general) is *not* symmetric: $\mathcal{D}_{\text{KL}}(p(\theta)\|q(\theta)) \neq \mathcal{D}_{\text{KL}}(q(\theta)\|p(\theta))$. The *symmetrized (KL-) divergence* is defined as

$$\mathcal{D}_{\text{s}}(p(\theta)\|q(\theta)) = \mathcal{D}_{\text{KL}}(p(\theta)\|q(\theta)) + \mathcal{D}_{\text{KL}}(q(\theta)\|p(\theta)) \quad (2)$$

(Kullback and Leibler 1951). Unlike the *directed* divergence, $\mathcal{D}_{\text{s}}$ is obviously symmetric. Note that, trivially but importantly,

$$\mathcal{D}_{\text{s}}(p(\theta)\|q(\theta)) \geq \max\{\mathcal{D}_{\text{KL}}(p(\theta)\|q(\theta)),\ \mathcal{D}_{\text{KL}}(q(\theta)\|p(\theta))\}, \quad (3)$$

that is, the symmetrized divergence bounds both individual directed divergences. For simplicity, in the following we will mostly be focusing on *symmetrized* KL-divergences.

For example, the Kullback–Leibler divergence for two normal distributions with mean and variance parameters (μ_A, σ_A^2) and (μ_B, σ_B^2), respectively, is given by

$$\mathcal{D}_{\text{KL}}(p(\theta|\mu_A, \sigma_A)\|p(\theta|\mu_B, \sigma_B)) = \frac{1}{2}\left(\frac{(\mu_A-\mu_B)^2}{\sigma_B^2} + \frac{\sigma_A^2}{\sigma_B^2} + \log\left(\frac{\sigma_B^2}{\sigma_A^2}\right) - 1\right) \quad (4)$$

CONTACT Christian Röver christian.roever@med.uni-goettingen.de Department of Medical Statistics, University Medical Center Göttingen, Göttingen 37073, Germany.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

(Kullback 1959, chap. 9). The symmetrized divergence then results as

$$\mathcal{D}_{\mathrm{s}}(p(\theta|\mu_A,\sigma_A)\|p(\theta|\mu_B,\sigma_B)) = \frac{(\mu_A-\mu_B)^2}{\left(\frac{1}{2}\left(\sigma_A^{-2}+\sigma_B^{-2}\right)\right)^{-1}} + \frac{\left(\sigma_A^2-\sigma_B^2\right)^2}{2\,\sigma_A^2\,\sigma_B^2}. \quad (5)$$

2.2. Motivation and Interpretation

The Kullback–Leibler divergence is generally regarded as a measure of *discrepancy* between probability distributions. For example, when a simple parametric approximation to a more complicated distribution is sought, the approximation may reasonably be matched against the true distribution via minimization of the divergence (Bernardo and Smith 1994; O'Hagan 1994).

The divergence $\mathcal{D}_{\mathrm{KL}}$ relates to the logarithmic ratio of densities. The domain of main interest here is the limit of very similar p and q, that is, almost equal numerator and denominator, when the density ratio is close to unity. In that case the logarithmic ratio approximately corresponds to the "relative difference" in densities: since $\log(x) \approx x-1$ for $x \approx 1$ (and hence $\log(a/b) \approx a/b - 1$ for $a \approx b$), a divergence of, say, 0.01 approximately corresponds to an (expected) 1% difference between numerator and denominator.

While there is no simple connection relating the divergence of two distributions to their moments, one can get an impression by considering the generic case of two normal distributions. For some fairly obvious parameter choices, we get

$$\sigma_B=\sigma_A,\quad \mu_B=\mu_A+c\sigma_A \;\Rightarrow\; \mathcal{D}_{\mathrm{KL}}(p\|q)=\tfrac{1}{2}c^2, \quad (6)$$

$$\mathcal{D}_{\mathrm{s}}(p\|q)=c^2, \quad (7)$$

and

$$\mu_B=\mu_A,\quad \sigma_B=(1+c)\sigma_A \;\Rightarrow\; \mathcal{D}_{\mathrm{KL}}(p\|q)=\frac{1}{2(1+c)^2} + \log(1+c)-\frac{1}{2} \approx c^2, \quad (8)$$

$$\mathcal{D}_{\mathrm{s}}(p\|q)=\frac{c^2(c+2)^2}{2(c+1)^2} \approx 2c^2, \quad (9)$$

where the latter approximations follow from Taylor expansion around $c=0$.

From the above, we can see that, for example, for equal variances, a difference in means by, say, 1% of a standard deviation corresponds to a symmetrized divergence $\mathcal{D}_{\mathrm{s}}=0.01^2=0.0001$. For equal means on the other hand, standard deviations differing by 1% correspond to a symmetrized divergence of ≈ 0.0002.

3. Mixture Distributions and Discrete Approximations

3.1. Definitions

Suppose a random variable Y follows a distribution with density $p(y|x)$ that depends on a parameter x. If that parameter is not fixed, but again is a random variable (X) with density $p(x)$, then the (marginal) distribution of Y is called a *mixture distribution*. The joint density of X and Y is given by $p(x,y)=p(y|x)\times p(x)$. What is commonly of interest is the marginal (unconditional) distribution of Y, whose density results by integration as $p(y)=\int p(x,y)dx=\int p(y|x)\,p(x)dx$. The (marginal) distribution of the underlying variable that is conditioned upon, $p(x)$, is called the *mixing distribution* (Seidel 2010) or *latent distribution* (Lindsay 1995).

Mixture distributions arise frequently in statistical problems, for example, as marginal (posterior) distributions or as convolutions of random variables. In the following, we will assume that X is one-dimensional, and that the domain of X is the real line, or a subset thereof (continuous or discrete).

3.2. Binning

To transition from continuous to discrete mixtures, we define a *binning* of the domain of X. Let $\{x_{(1)}, x_{(2)}, \ldots, x_{(k-1)}\} \subset \mathbb{R}$ be a set of bin margins with $x_{(1)} < x_{(2)} < \cdots < x_{(k-1)}$. These define the (exhaustive and disjoint) set of k bins $\{\mathcal{X}_i\}_{i=1,\ldots,k}$ with

$$\mathcal{X}_i = \begin{cases} \{x : x \leq x_{(1)}\} & \text{if } i=1 \\ \{x : x_{(i-1)} < x \leq x_{(i)}\} & \text{if } 1<i<k \\ \{x : x_{(k-1)} < x\} & \text{if } i=k. \end{cases} \quad (10)$$

In addition, the set of k points $\{\tilde{x}_1,\ldots,\tilde{x}_k\}$ with $\tilde{x}_i \in \mathcal{X}_i$ defines a set of *reference points*, one for each bin. Each bin also has a probability π_i (with respect to $p(x)$) associated, which is given by

$$\pi_i = \mathrm{P}\big(x_{(i-1)} < x \leq x_{(i)}\big) = \mathrm{P}\big(x \in \mathcal{X}_i\big). \quad (11)$$

3.3. The Binned Mixture

In addition to the probability density $p(x,y)$ given above, we define another probability distribution with density $q(x,y)$ that has the same marginal density (mixing distribution)

$$q(x) = p(x), \quad (12)$$

and whose conditional probability density is given by

$$q(y|x) = p(y|x{=}\tilde{x}_i) \quad \text{for } x \in \mathcal{X}_i. \quad (13)$$

So q is similar to p, but instead of conditioning on the "exact" x value as in the original definition above, this probability distribution conditions on the corresponding bin's reference value $\tilde{x}_i$, depending on which bin x belongs to. The joint distribution of X and Y again is defined through its joint density: $q(x,y)=q(x)\times q(y|x)$. The marginal density of Y again turns out as $q(y)=\int q(y|x)\,q(x)dx$. Equivalently, the binning may be considered a discretization of the mixing distribution while keeping the conditional distribution the same. The discretized mixing distribution simply has the reference points $\{\tilde{x}_1,\ldots,\tilde{x}_k\}$ as its domain, while the associated bin probabilities $\{\pi_1,\ldots,\pi_k\}$ define the probability mass function. The reference points consequently act as "support points" for the discretized mixing distribution here; alternating between these points of view is sometimes helpful.

This "binned" approximation to the joint distribution of (X,Y) is useful, as the resulting marginal distribution of Y, $q(y)$, is a discrete sum of conditional densities (rather than an integral), making numerical evaluation very easy. The marginal density simplifies to

$$q(y) \;=\; \sum_{i=1}^{k} \pi_i \, p(y|\tilde{x}_i). \tag{14}$$

Analogously, the cumulative distribution function (CDF) may also be expressed as a weighted sum of the component CDFs. Random number generation as well as computation of moments for finite mixtures is also straightforward (Lindsay 1995).

4. Constructing Binned Mixture Approximations

4.1. Some Preliminary Results

For each bin i define the maximum symmetrized KL-divergence

$$d_i \;=\; \max_{x\in\mathcal{X}_i}\{\mathcal{D}_{\mathrm{s}}(p(y|x)\|p(y|\tilde{x}_i))\} \;=\; \max_{x\in\mathcal{X}_i}\{\mathcal{D}_{\mathrm{s}}(p(y|x)\|q(y|x))\}, \tag{15}$$

that is, the maximum (symmetrized) divergence between distributions $p(y|x)$ corresponding to points within the ith bin and the corresponding ith reference point.

The *chain rule for relative entropy* states that

$$\mathcal{D}_{\mathrm{KL}}(p(x,y)\|q(x,y)) = \mathcal{D}_{\mathrm{KL}}(p(x)\|q(x)) + \mathrm{E}_{p(x)}[\mathcal{D}_{\mathrm{KL}}(p(y|x)\|q(y|x))] \tag{16}$$

(Cover and Thomas 1991, sec. 2.5). In other words, the divergence of two joint distributions is the sum of the divergence of the marginals and the expected divergence of the conditionals. Note that the expectation in (16) is also known as the *conditional relative entropy* (Cover and Thomas 1991, sec. 2.5). For the *symmetrized* divergence immediately follows an analogous property:

$$\mathcal{D}_{\mathrm{s}}(p(x,y)\|q(x,y)) = \mathcal{D}_{\mathrm{s}}(p(x)\|q(x)) + \mathrm{E}_{p(x)}[\mathcal{D}_{\mathrm{s}}(p(y|x)\|q(y|x))]. \tag{17}$$

In our case, we have identical marginal distributions for X under both distributions, $p(x)=q(x)$, so that

$$\mathcal{D}_{\mathrm{KL}}(p(x)\|q(x)) \;=\; \mathcal{D}_{\mathrm{KL}}(q(x)\|p(x)) \;=\; 0 \tag{18}$$

and consequently

$$\mathcal{D}_{\mathrm{s}}(p(x,y)\|q(x,y)) = \mathrm{E}_{p(x)}[\mathcal{D}_{\mathrm{s}}(p(y|x)\|q(y|x))]. \tag{19}$$

We are interested in the approximation of $p(x,y)$ through the simplified distribution $q(x,y)$, and in particular of $p(y)$ by $q(y)$. We know, again via the chain rule, that

$$\begin{aligned}
\mathcal{D}_{\mathrm{s}}(p(y)\|q(y)) &\overset{(17)}{=} \mathcal{D}_{\mathrm{s}}(p(x,y)\|q(x,y)) \\
&\quad - \mathrm{E}_{p(x)}[\mathcal{D}_{\mathrm{s}}(p(x|y)\|q(x|y))] && (20)\\
&\leq \mathcal{D}_{\mathrm{s}}(p(x,y)\|q(x,y)) && (21)\\
&\overset{(18)}{=} \mathrm{E}_{p(x)}[\mathcal{D}_{\mathrm{s}}\big(p(y|x)\|q(y|x)\big)] && (22)\\
&\leq \sum_i \pi_i d_i && (23)\\
&\leq \max_i d_i \;=:\; \delta. && (24)
\end{aligned}$$

So, by limiting the divergences of conditionals $p(y|x)$ and $q(y|x)$ within each single bin such that these remain $\leq \delta$ (24), we can now also bound the divergence of exact and approximate marginals $p(y)$ and $q(y)$ (20).

4.2. The Proposed Approach

Given the bin-wise divergences, we can now bound the divergence of exact and discretized marginals. The obvious question now is whether and how one can invert the argument and construct a grid approximation matching a prespecified maximum divergence δ. For a given (reference) point x in the mixing distribution's domain, we can find a corresponding neighborhood within which the divergence remains below δ. Once we have defined a single bin this way, we can also generate an exhaustive covering of the whole parameter space through such bins. We abbreviate this method as the DIRECT *(divergence restricting conditional tesselation) approach*, as it aims at a covering of the conditional's parameter space while bounding the divergence.

In some cases it is not possible to have a finite number of bins associated with finite bin-wise divergences. A "trick," if necessary, then is to simply ignore some fraction of parameter space (of the mixing distribution's domain) that is associated with a preset, arbitrarily small probability ϵ and do the binning on the remaining share of parameter space. Problems with unbounded divergences, or infinite numbers of necessary bins, commonly occur toward one or both of the parameter space's margins. Neglecting a certain fraction of parameter space that is associated with an (arbitrarily) small probability ϵ will usually not pose a significant practical problem, as it will only add another bit to the error budget that needs to be considered in (almost) any numerical computation anyway.

4.3. The Sequential DIRECT Algorithm

We will in the following construct a binning so that the resulting discrete approximation of the exact marginal does not differ, in terms of symmetrized divergence, and with that of both directed divergences, from the exact ("continuous") marginal by more than a prespecified amount. The number (k) of components and the placement of reference points will be determined automatically in the process. The idea is to sequentially divide the mixing distribution's domain into bins, while first ensuring that the divergences within bins are bounded, and second, if necessary, ignoring the mixing distribution's extreme left and/or right tails. To proceed, in the following we will assume that the divergence between any pair of points (x_1, x_2) in parameter space is Lipschitz continuous, at least within a range $[\tilde{x}_1, \tilde{x}_k]$ with $\mathrm{P}\big(X \notin [\tilde{x}_1, \tilde{x}_k]\big) \leq \epsilon$. This will ensure that the algorithm will work, although violations do not necessarily prevent a solution; even continuity is not strictly necessary. A possible implementation of the DIRECT approach is defined in Table 1.

Table 1. The sequential DIRECT algorithm (see Section 4.3).

1. Specify a maximum KL-divergence $\delta > 0$, some small probability $0 \leq \epsilon \ll 1$, and a starting reference point $\tilde{x}_1$. Sensible values for $\tilde{x}_1$ may, for example, be the minimum possible value, the $\frac{\epsilon}{2}$-quantile, or any value with $\mathrm{P}(X \leq \tilde{x}_1) < \epsilon$. Define $\epsilon_1 := \mathrm{P}(X \leq \tilde{x}_1) \geq 0$. Set $i = 1$.
2. Set $x^\star = \tilde{x}_1$. Obviously, $\mathcal{D}_s(p(y|\tilde{x}_1)\|p(y|x^\star)) = 0$. Now increase $x^\star$ as far as possible while ensuring that $\mathcal{D}_s(p(y|\tilde{x}_1)\|p(y|x^\star)) \leq \delta$. Use this point as the first bin margin: $x_{(1)} = x^\star$. Compute $\pi_1 = \mathrm{P}(x < x_{(1)})$. Set $i = i + 1$.
3. Increase $x^\star$ until $\mathcal{D}_s(p(y|x_{(i-1)})\|p(y|x^\star)) = \delta$. Use this point as the next reference point: $\tilde{x}_i = x^\star$.
4. Increase $x^\star$ again until $\mathcal{D}_s(p(y|\tilde{x}_i)\|p(y|x^\star)) = \delta$. Use this point as the next bin margin: $x_{(i)} = x^\star$.
5. Compute the bin weight $\pi_i = \mathrm{P}(x_{(i-1)} < X \leq x_{(i)})$.
6. If $\mathrm{P}(X > x_{(i)}) > (\epsilon - \epsilon_1)$, set $i = i + 1$ and proceed at step 3. Otherwise stop.

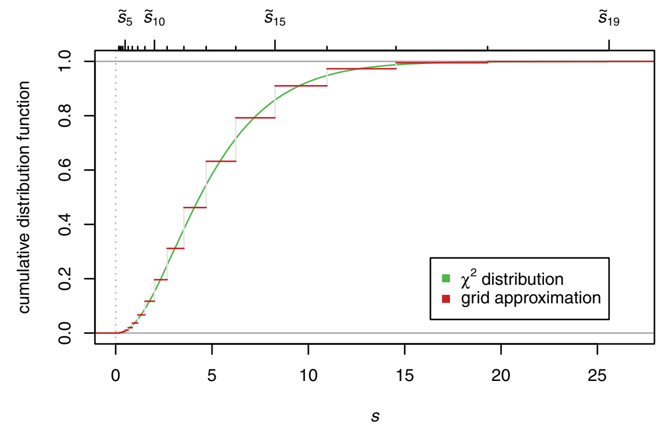

Figure 1. The underlying χ^2 mixing distribution (of the latent variable s) and the grid approximation that is effectively used instead in the Student-t example (Section 5.1). The extra tick marks at the top indicate the 19 grid points used.

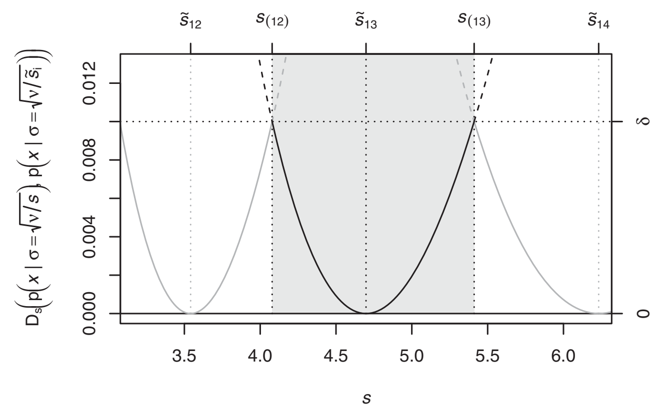

Figure 2. Illustration of how the binning is set up (Student-t example, Section 5.1). Bin margins $s_{(i)}$ and reference points $\tilde{s}_i$ are arranged such that within each bin the divergence relative to the corresponding reference point does not exceed the preset threshold δ.

Reference points $\tilde{x}_i$ and corresponding weights π_i now allow to define an approximation q as in (14). It is actually not necessary to also keep track of the exact bin margins $x_{(i)}$ once the bin weights π_i are determined. The maximum divergence of conditionals, and with that of the marginals, will now be $= \delta$, possibly up to a bit of probability ($\leq \epsilon$) beyond the first and/or last bins.

The essence here is to ensure condition (24) to be met. Possible boundary or singularity problems are circumvented by ignoring negligible bits of parameter space via specification of ϵ. Lipschitz continuity of the divergence will ensure that the relevant range may be covered using a finite number of bins. Note that the actual form of the latent (mixing) distribution is only used to determine the relevant range in parameter space, while the actual binning is otherwise independent. A number of variations of the DIRECT algorithm are conceivable, for example, it may or may not be sensible, or possible, to either have a reference point or a bin margin at the parameter space's boundary. Also, the relationship between x and $p(y|x)$ may not necessarily be monotonic, in which case it may be possible to devise more efficient nonsequential binning strategies.

5. Examples

5.1. Student-t Distribution

A prominent example of a mixture distribution is the Student-t distribution. It arises as a continuous mixture of normal distributions with zero mean and scale $\sigma = \sqrt{\frac{\nu}{s}}$, where s is a draw from a χ^2 distribution with ν degrees of freedom (Johnson, Kotz, and Balakrishnan 1994, chap. 28). We can approximate the marginal Student-t distribution by a mixture of normal distributions, conditioning on a finite set of grid points in s, and compare against the true marginal, which in this case we know to be a Student-t distribution.

Suppose we are interested in the case of $\nu = 5$ degrees of freedom. We set the tuning parameters to $\delta := 0.01$ and $\epsilon := 0.001$ and we use the χ^2_5 distribution's $\frac{\epsilon}{2}$-quantile as the starting reference point ($\tilde{s}_1 := 0.158$). Applying the sequential DIRECT algorithm from Section 4.3 (using expression (5)) results in a set of 19 reference points $\tilde{s}_i$. As a result from the implied differences in the corresponding conditionally normal distributions, the 19 reference points are very unequally spaced, with many points concentrated near zero and a coarser spacing at large values (see Figure 1).

Figure 2 illustrates the construction of the binning by showing the 13th bin and its two neighboring bins with bin margins $s_{(i)}$ and reference points $\tilde{s}_i$. One can see that by construction within each bin the divergence relative to the corresponding reference point, $\mathcal{D}_s(p(x|\sigma = \sqrt{\nu/s}),\ p(x|\sigma = \sqrt{\nu/\tilde{s}_i}))$, remains below δ.

The 19-component normal mixture approximation is compared to the true marginal distribution in Figure 3. The two densities are barely distinguishable, and their ratio is very close to unity; it only diverges toward the distributions' extreme tails. The numerically computed actual divergence in this case amounts to $\mathcal{D}_s\big(p(x)||q(x)\big) \approx 3.5 \times 10^{-5}$.

5.2. Convolution of Two Distributions

In the following, we present the example of computing the convolution of two distributions. Suppose we have two random variables, X and Y, with densities $p_X(x)$ and $p_Y(y)$. We are interested in their sum $Z = X + Y$, and its density $p_Z(z)$. Here, we take X and Y to follow skew-normal and logistic distributions, respectively, so that the solution is not trivial. We can turn the problem into that of a mixture distribution and subsequently apply the above algorithm by first considering the joint distribution of X and Z. Note that $\mathrm{P}(Z = z \,|\, X = x) = \mathrm{P}(Y =$

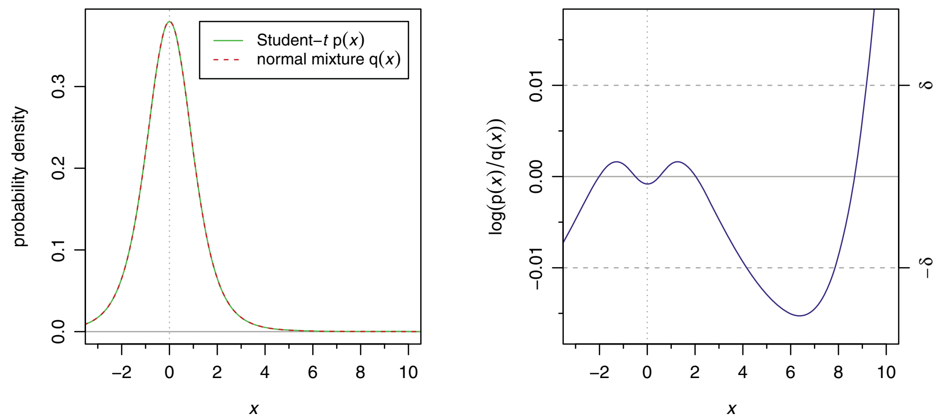

Figure 3. Comparison of the true mixture distribution, the Student-*t* distribution, to the grid approximation. The left panel shows the two probability density functions on top of each other; the two are essentially undiscernible at this scale. The right panel shows the logarithmic ratio of the densities as a function of *x*.

$z - x$), so the conditional distribution of $Z|X$ here is simply a "shifted" version of the (known) distribution P_Y. With that, we can rewrite the target density p_Z as a marginal density in terms of the (known) marginal $p_X(x)$ and the (known) conditional $p_Z(z|x) = p_Y(z - x)$:

$$p_Z(z) \;=\; \int p_Y(z-x)\, p_X(x) dx. \qquad (25)$$

This way it is obvious that convolution of two random variables may again be seen as a special case of a mixture distribution where the conditional $\mathrm{P}(Z|X)$ is mixed via the latent distribution $\mathrm{P}(X)$. Due to symmetry of the problem, the roles of X and Y may also be reversed.

In the following, suppose that $p_X(x) > 0$ and $p_Y(y) > 0$ for all $x, y \in \mathbb{R}$, that is, the domain of both X and Y is the whole real line. When applying the DIRECT algorithm to set up an approximation, it is important to note that the divergence $\mathcal{D}_s\big(p_Z(z|x_1)\,\|\,p_Z(z|x_2)\big)$ required in steps 2–4 of the algorithm (Section 4.3) only depends on the (absolute) difference $|x_2 - x_1|$, since the conditional distributions $p_Z(z|\cdot)$ here only differ by a shift in location. This implies that the bin width $\tilde{x}_i - \tilde{x}_{i-1}$ is constant across all bins, and hence only needs to be determined once. This simplifies the grid construction to a few steps:

1. Determine the bin half-width Δ_x such that $\mathcal{D}_s(p_Y(y)\|p_Y(y - \Delta_x)) = \delta$.
2. Determine minimum and maximum X values $\tilde{x}_1$ and $\tilde{x}_k$, for example, as the $\frac{\epsilon}{2}$ and $1-\frac{\epsilon}{2}$ quantiles of p_X.
3. Determine the remaining reference points $\tilde{x}_2$ to $\tilde{x}_{k-1}$ as well as their total number k by filling the interval with reference points that are at most $(\tilde{x}_i - \tilde{x}_{i-1}) \leq \Delta_x$ apart.

A general implementation of the procedure in R is shown in the online supplement. Divergences here are computed numerically, without needing to have the corresponding formulas available in analytic form.

Consider the example of the sum of two random variables, one following a skew-normal distribution with shape parameter $\alpha = 4$ (Azzalini 2014, 2015), and one following a logistic distribution. Application of the DIRECT algorithm (using $\delta = 0.01$ and $\epsilon = 0.001$) results in a 13-component mixture of logistic distributions to approximate the convolution. Draws from the two summands' distributions may easily be simulated, so it is straightforward to also generate samples of their sum's distribution. Figure 4 illustrates the fit of the numerical approximation to 1,000,000 simulated samples via a quantile-quantile plot (Q-Q plot). Here the 10 smallest and largest samples are shown as individual dots, other quantiles are connected by a line, and selected quantiles are highlighted. Note that while the design parameter ϵ was set to 0.001, the simulated and computed quantiles appear to match well even beyond tail probabilities of 0.001. The R code to reproduce these simulations is also provided in the online supplement. All computations here were carried out using R (R Core Team 2015).

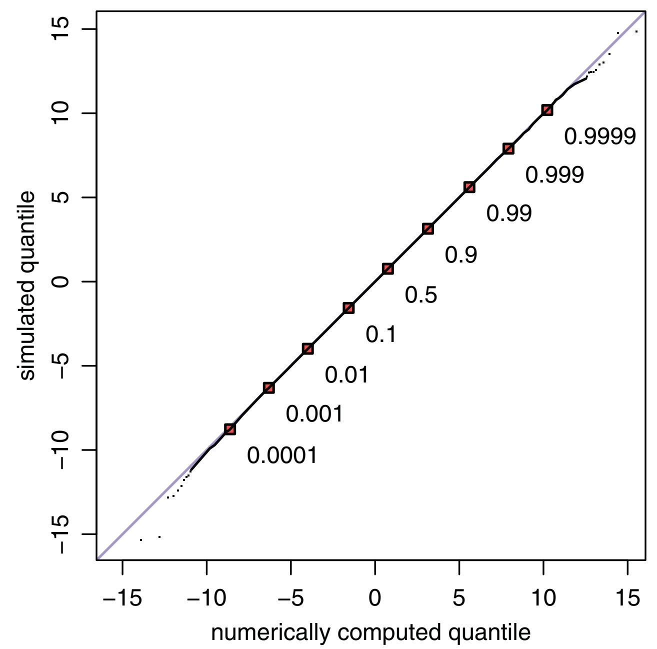

Figure 4. Q-Q-plot illustrating the accuracy of the convolution of a skew-normal and a logistic distribution by comparing quantiles computed numerically, using the DIRECT algorithm, and simulated quantiles.

6. Conclusions

The DIRECT approach introduced in this article allows to generate finite mixtures as approximations to mixture distributions with a large or infinite number of mixture components. A formulation in terms of a finite mixture distribution then makes density function, cumulative distribution function, etc., easily accessible. The mismatch incurred by resorting to the approximation is efficiently controlled via two tuning parameters (δ and ϵ). The described algorithm allows for easy implementation in a completely automated fashion, as is also demonstrated in the examples. The setup relies on the computation of (symmetrized) divergences of (conditional) distributions; ideally these are available analytically, but numerical computation is also not a problem.

Variations of the DIRECT algorithm are conceivable. The bound derived in Section 4 may be met in many different ways; the described one is only a simple, general solution. For example, it may be possible, and possibly more efficient, to aim at the condition in (23) rather than (24) to bound the divergence. While for simplicity we concentrated on symmetrized divergences here, it may also make sense to directly aim for directed Kullback–Leibler divergences instead.

A generalization to higher dimensions of the latent mixing distribution should in general also be possible. Since the problem of covering of higher-dimensional spaces is considerably trickier, it may eventually be easiest to resort to random coverings here (Messenger, Prix, and Papa 2009; Röver 2010).

The algorithm was originally developed and eventually applied in the context of the `bayesmeta` R package (Röver 2015). In this meta-analysis application, one is faced with the common problem of inferring two parameters (τ and μ) via their posterior probability distribution. From their joint distribution ($p(\mu, \tau)$), one of the marginals, $p(\tau)$, may be derived analytically, while the conditionals $p(\mu|\tau)$ are normal. Primary interest usually lies in μ, and application of the DIRECT algorithm facilitates quick and accurate computation of the marginal $p(\mu)$ without having to use, for example, Markov chain Monte Carlo (MCMC) methods (Friede et al. 2016).

Supplementary Materials

The supplementary material includes two R code files that allow to reproduce the two examples discussed in Section 5.

Acknowledgments

This project has received funding from the European Union's *Seventh Framework Programme for research, technological development and demonstration* under grant agreement number FP HEALTH 2013-602144 "*Innovative methodology for small populations research (InSPiRe).*"

ORCID

Christian Röver http://orcid.org/0000-0002-6911-698X
Tim Friede http://orcid.org/0000-0001-5347-7441

References

Azzalini, A. (2014), *The Skew-Normal and Related Families* (*Institute of Mathematical Statistics Monographs*), Cambridge: Cambridge University Press. [221]

—— (2015), "The Skew-Normal Probability Distribution," available at *http://azzalini.stat.unipd.it/SN* [221]

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, Chichester, UK: Wiley. [218]

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory*, New York: Wiley. [217,219]

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2016), "Meta-Analysis of Few Small Studies in Orphan Diseases," *Research Synthesis Methods*. Available at *http://arxiv.org/abs/1601.06533*. [222]

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions* (2nd ed.), New York: Wiley. [220]

Kullback, S. (1959), *Information Theory and Statistics*, New York: Wiley. [218]

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86. [217]

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications* (*NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5), Hayward, CA: Institute of Mathematical Statistics. [218,219]

Messenger, C., Prix, R., and Papa, M. A. (2009), "Random Template Banks and Relaxed Lattice Coverings," *Physical Review D*, 79, 104017. [222]

O'Hagan, A. (1994), *Bayesian Inference* (*Kendall's Advanced Theory of Statistics*, Vol. 2B), New York: Wiley. [218]

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *https://www.r-project.org/* [221]

Röver, C. (2010), "Random Template Placement and Prior Information," *Journal of Physics: Conference Series*, 228, 012008. [222]

—— (2015), "bayesmeta: Bayesian Random-Effects Meta Analysis," R package. Available at *http://cran.r-project.org/package=bayesmeta* [222]

Seidel, W. E. (2010), "Mixture Models," in *International Encyclopedia of Statistical Science*, ed. M. Lovric, Heidelberg: Springer, pp. 827–829. [217,218]