Finite Gaussian Mixture Approximations to Analytically Intractable Density Kernels

Khorunzhina, Natalia; Richard, Jean-Francois

Document Version
Accepted author manuscript

Published in:
Computational Economics

DOI:
10.1007/s10614-017-9777-2

Publication date:
2019

License
Unspecified

Citation for published version (APA):
Khorunzhina, N., & Richard, J.-F. (2019). Finite Gaussian Mixture Approximations to Analytically Intractable Density Kernels. *Computational Economics*, *53*(3), 991-1017. https://doi.org/10.1007/s10614-017-9777-2

Finite Gaussian Mixture Approximations to Analytically Intractable Density Kernels

Natalia Khorunzhina and Jean-Francois Richard

Journal article (Accepted version*)

Please cite this article as:
Khorunzhina, N., & Richard, J-F. (2019). Finite Gaussian Mixture Approximations to Analytically Intractable Density Kernels. Computational Economics, 53(3), 991-1017. https://doi.org/10.1007/s10614-017-9777-2

This is a post-peer-review, pre-copyedit version of an article published in Computational Economics. The final authenticated version is available online at:

DOI: https://doi.org/10.1007/s10614-017-9777-2

Noname manuscript No.
(will be inserted by the editor)

Finite Gaussian Mixture Approximations to Analytically Intractable Density Kernels

Natalia Khorunzhina · Jean-François Richard

Received: date / Accepted: date

Abstract The objective of the paper is that of constructing finite Gaussian mixture approximations to analytically intractable density kernels. The proposed method is adaptive in that terms are added one at the time and the mixture is fully re-optimized at each step using a distance measure that approximates the corresponding importance sampling variance. All functions of interest are evaluated under Gaussian product rules. Since product rules suffer from an obvious curse of dimensionality, the proposed algorithm as presented is only applicable to models whose non-linear and/or non-Gaussian subspace is of dimension up to three. Extensions to higher-dimensional applications would require the use of sparse grids, as discussed in the paper. Examples include a sequential (filtering) evaluation of the likelihood function of a stochastic volatility model where all relevant densities (filtering, predictive and likelihood) are closely approximated by mixtures.

Jean-François Richard acknowledges support from the National Science Foundation under grant no.1529151.

Natalia Khorunzhina
Copenhagen Business School
Department of Economics
Porcelænshaven 16A
2000 Frederiksberg, Denmark
Tel.: +45 3815 2403
E-mail: nk.eco@cbs.dk

Jean-François Richard
University of Pittsburgh
Department of Economics
4917 Wesley W. Posvar Hall
Pittsburgh, PA 15260, USA
Tel.: 1 412 6481750
E-mail: fantin@pitt.edu

Keywords Finite mixture · Distance measure · Gaussian quadrature · Importance sampling · Adaptive algorithm · Stochastic volatility · Density kernel

1 Introduction

Starting with early contributions more than a century ago by Newcomb (1886), Holmes (1892), Weldon (1892, 1893), and Pearson (1894) among others, finite mixtures have been continuously used in statistics (see section 2.18 in McLachlan and Peel 2000 for a short history of finite mixture models and Table 2.13 in Titterington et al 1985 for an extensive list of direct applications of mixtures; see also the monographs of Everitt and Hand 1981, Scott 1992, and Frühwirth-Schnatter 2006). More recently, mixtures of normal distributions have been increasingly applied in macro- and micro-economics (e.g., regime-switching models of economic time series in Hamilton 1989, or analysis of dynamics of educational attainment in Keane and Wolpin 1997, and Cameron and Heckman 2001), marketing science (structured representation of market information in DeSarbo et al 2001, and forecasting of new product sales in Moe and Fader 2002), and empirical finance (modeling stock returns in Kon 1984, and Tucker 1992, value-at-risk in Duffie and Pan 1997, Venkataraman 1997, and Hull and White 1998, stochastic volatility models in Kim et al 1998 and Omori et al 2007).

In the present paper we focus our attention on the specific problem of using finite mixture of Gaussian densities for approximating a non-standard density kernel. Such approximations are critically needed when inference requires numerical integration of an analytically intractable density kernel, such as a marginal likelihood for a non-linear and/or non-Gaussian state-space model or a Bayesian posterior density . Whether one relies upon direct numerical integration (Gaussian quadratures) or simulation methods such as Importance Sampling (IS) or Markov Chain Monte Carlo (MCMC), the numerical accuracy of the results critically depends on the quality of approximation. For example, an inefficient inportance sampler might require prohibitive number of draws to produce accurate results, or might even fail to converge due to tail problems (see, e.g., Geweke, 1996)

Finite mixtures are conceptually attractive within this context since theoretically they can produce accurate approximations to most density functions, depending upon the number of components (Ferguson, 1973).

There exist a vast literature which proposes various procedures for constructing finite (mostly Gaussian) mixture approximations. In a nutshell, the key numerical issues are the selection of a distance measure to assess goodness of fit, the (typically sequential) determination of the number of terms in the approximating mixtures and the estimation of its component parameters and weights.

Extending earlier proposals by West (1992), Oh and Berger (1993), Cappé, Guillin, Marin, and Robert (2004), and Douc, Guillin, Marin, and Robert

(2007), Cappé, Douc, Guillin, Marin, and Robert (2008) proposes an adaptive algorithm to optimize the IS performance of a mixture sampler with a predetermined number of components. Specifically, their Mixture Population Monte Carlo (M-PMC) algorithm aims at maximizing the entropy criterion between a target kernel and the mixture approximation. It is adaptive in that it relies upon sampling from the current mixture proposal in updating its weights and component parameters. Convergence is assessed on the basis of the Shannon entropy of the normalized IS ratios.

Hoogerheide et al (2007) propose an adaptive algorithm to construct mixtures of Student-t distributions to approximate an arbitrary target density with the objective of minimizing the variance of the corresponding IS ratios. Adaption means that the components of the mixture are introduced sequentially until a good enough fit is obtained. This algorithm has been implemented within the R package AdMit in Ardia et al (2009). A subsequent adaptive algorithm is developed by Hoogerheide et al (2012) and implemented into the R package MitISEM by Basturk et al (2012). As we shall see, the algorithm we propose below is adaptive in the sense of Basturk et al (2012), but differs in several important ways: it relies upon a different distance measure; the latter is evaluated by Gaussian quadrature instead of importance sampling (classical) or Metropolis-Hastings (bayesian); optimization relies upon an analytical gradient optimizer and initial values are computed differently.

Giordani and Kohn (2010) propose an adaptive Independent Metropolis-Hastings algorithm for constructing mixture proposal densities. Fast re-estimation of the mixtures relies upon a k-means algorithm discussed in Bradley and Fayyad (1998) and subsequently in Hamerly and Elkan (2002) and Giordani and Kohn (2010). Efficient designs rely upon reducing the number of re-estimations as coverage improves.

Kurtz and Song (2013) propose a Cross-Entropy-Based Adaptive Importance Sampling algorithm to construct an optimal Gaussian mixture IS density with a preassigned number of terms. The objective function that is sequentially minimized is the Kullback-Leibler cross-entropy between the target density and the mixture.

The approach of Bornkamp (2011) relies upon iterated Laplace approximations to add components one by one as needed. However, only the weights of the mixture components are re-optimized with each iteration while their Laplace modes and inverted Hessians are left unchanged. It immediately follows that a mixture target cannot be reproduced. In sharp contrast our algorithm includes full sequential re-optimization to the effect that if the target density is a mixture it reproduces it exactly as we shall illustrate in section 3.1 for example 2 in Bornkamp (2011).

In this paper we propose a fully adaptive algorithm to construct Gaussian mixture approximations to a low-dimensional ($n \leq 3$) target density kernel. Our algorithm is also applicable in higher dimensional models that can be factorized into a linear Gaussian conditional density and a marginal non-standard density to be approximated by a mixture. An example of such dimension reduction is provided in section 3.2 below. Our algorithm includes

full re-optimization with the introduction of each additional component. Since such mixture approximations will often be used as importance sampling or proposal densities, we use an efficient importance sampling (EIS) approximation of the sampling variance as our distance measure to be minimized, whereby optimization takes the form of an auxiliary non-linear least squares problem.

Our algorithm is illustrated by several test cases. The first application approximates a mixture of three bivariate normal distributions and demonstrates the ability of the proposed algorithm to exactly reproduce the target mixture. The second application approximates a bivariate skew-distribution, a class of densities of growing importance in economics (modeling fertility patterns in Mazzuco and Scarpa, 2015, stochastic frontier analysis in Domínguez-Molina et al, 2004, sample selection models in Marchenko and Genton, 2012; Ogundimu and Hutton, 2016) and finance (capital asset pricing models in Adcock, 2004, 2010). Our third application deals with a basic stochastic volatility model, whose measurement density can be approximated by a mixture of normal distributions (see, e.g. Kim et al, 1998; Omori et al, 2007). The potential scope of applications of our procedure is not limited to approximating analytically intractable densities. Our procedure provides alternative numerical solutions to a wide range of problems in economics and finance, some of which we outline in the paper.

The paper is organized as follows: the baseline algorithm is presented in section 2; examples are presented in section 3. In section 4, we discuss future research plans together with pilot applications. Section 5 concludes. Technical derivations are regrouped in as Appendix.

2 Mixture approximation

2.1 Notation

Let $\varphi(x)$ denote the target (density) kernel to be approximated. Its integrating constant on the support $D \subset \mathbb{R}^d$ is given by

$$G = \int_D \varphi(x)dx \tag{1}$$

and is typically unknown. We note that φ and G could depend on unknown parameters in which case the approximations presented below would have to be re-computed for each new parameter value. Dependence on such parameters is omitted in our notation for ease of presentation. Let $k(x, \alpha)$ denote a parametric Gaussian kernel of the form

$$k(x, \alpha) = |R| \exp\left[-\frac{1}{2}(x-\mu)^{'} RR^{'}(x-\mu)\right], \tag{2}$$

with R (Cholesky) lower triangular (with the elements r_{ij}, where $r_{ii} > 0$) and $\alpha = (\mu, R)$. Since G is generally unknown and not equal to 1, we aim at

constructing an un-normalized Gaussian mixture kernel of the form

$$k_J(x, a_J) = \sum_{j=1}^{J} e^{\delta_j} k(x, \alpha_j) \tag{3}$$

with $a_J = ((\alpha_1, \delta_1), ..., (\alpha_J, \delta_J))$. The corresponding importance sampling density is given by[1]

$$m_J(x|a_J) = \chi_J^{-1}(a_J) k_J(x, a_J) \tag{4}$$

$$\chi_J(a_J) = (2\pi)^{d/2} \sum_{j=1}^{J} e^{\delta_j} \tag{5}$$

with component probabilities

$$\pi_i = e^{\delta_i} \left(\sum_{j=1}^{J} e^{\delta_j} \right)^{-1}. \tag{6}$$

The corresponding IS ratios are proportional to

$$\nu(x, a_J) = \frac{\varphi(x)}{k_J(x, a_J)} \tag{7}$$

with proportionately constant $G^{-1}\chi_J(a_J)$.

2.2 Distance measure

Most of the approximation methods we have surveyed, as well as the one we propose, can be subsumed under the heading "minimum distance estimators". Table 4.5.1 in Titterington et al (1985) lists several distance measures that have been used in the literature and discusses their relative merits, noting that the choice of a distance measure can be very important and should, therefore, be guided by the intended usage of the approximations. Since most of the applications that we have in mind require the construction of efficient proposal densities for IS and MCMC, we rely upon the distance measure proposed by Richard and Zhang (2007) for EIS. It consists of a second order approximation to the sampling variance of the IS ratios in Equation (7) and is proportional to

$$f_J(a_J) = \frac{1}{2} \int_D \left[\ln \varphi(x) - \ln k_J(x, a_J) \right]^2 \varphi(x) dx. \tag{8}$$

Note the absence of an intercept in the squared difference. Inclusion of an intercept would indeed require that the mixture weights e^{δ_j} add up to 1 for identification. It is far more convenient to leave these weights unconstrained by setting the intercept equal to zero. This being said, in order to avoid potentially large imbalances between $\ln \varphi(x)$ and $\ln k_J(x, a_J)$, it is often advisable

[1] Or a truncated version thereof is D is a strict subset of $\mathbb{R}^d$.

to normalize $\varphi(x)$ by $(2\pi)^{d/2}\hat{G}_0$, where $\hat{G}_0$ denotes an initial estimate of G as obtained below. In such a case we might expect the sum of the mixture weights to get closer to 1 as J increases.

2.3 Gaussian integration

Obviously, $f_J(a_J)$ in Equation (8) has to be evaluated numerically. In order to apply IS for that purpose, Richard and Zhang (2007) propose replacing $f_J(a_J)$ in Equation (8) by

$$\tilde{f}_J(a_J) = \frac{1}{2}\int_D \left[\ln\varphi(x) - \ln k_J(x, a_J)\right]^2 m_J(x|a_J)dx. \tag{9}$$

While $\tilde{f}$ is not equivalent to f (unless $m_J(x|a_J)$ were proportional to $\varphi(x)$, in which case the problem is solved), it provides an alternative operational distance measure to approximate $\ln\varphi(x)$. Foremost, its IS estimate is then given by

$$\hat{f}_J(a_J) = \frac{1}{2S}\sum_{i=1}^{S}\left[\ln\varphi(\tilde{x}_i) - \ln k_J(\tilde{x}_i, a_J)\right]^2, \tag{10}$$

where $\{\tilde{x}_i\}_{i=1}^S$ denotes S i.i.d. draws from $m_J(x|a_J)$. Since these draws depend on a_J, minimization of $\hat{f}_J(a_J)$ obtains from a fixed point sequence whereby $\hat{a}_J^{[l]}$ is computed under draws from $m_J(x|\hat{a}_J^{[l]})$, with an initial estimate $\hat{a}_J^{[o]}$ obtained e.g. from Laplace approximations (see Richard and Zhang, 2007, for implementation details). However, we found out from initial trial runs that such a fixed point procedure cannot be recommended for mixtures since it fails to produce enough draws for reliable estimation of low probability mixture components (since, in particular, the gradient for α_j is proportional to e^{δ_j}, as discussed further in section 3 below).

Instead we propose to evaluate $f_J(a_J)$ using a product of univariate Gaussian quadrature. Product rules remain manageable for low dimensions, say $d \leqslant 3$. Higher dimensions require the use of sparse grids, as will be discussed in section 4. We can also take advantage of situations where $\varphi(x)$ can be partitioned into

$$\varphi(x) = \varphi_1(x_1)\varphi_2(x_2|x_1) \tag{11}$$

with x_1 low-dimensional and φ_2 a linear Gaussian kernel, in which case only φ_1 needs to be approximated by a mixture.

We implemented three different product rules based on Legendre, Hermite and Mixture-Hermite quadratures, all of which are paired with appropriate linear transformations of x. The key trade-off between Legendre and Hermite rules is largely depending on the tail behaviour of the target kernel. Hermite rules operate on $\mathbb{R}^d$ and will reach far in the tails of the target, but will often waste nodes (especially for product rules) in distant regions where tails become negligible for practical purposes. Legendre rules avoid that tail problem to a

large extent by operating on bounded subspaces of $\mathbb{R}^d$ but could fail to adequately capture tail behaviour in the case of excessive truncation (a problem that is, nevertheless, easy to detect by an additional trial run under increased range). An explicit comparison between the three rules is provided in section 3.2.

2.3.1 Legendre

Depending on how far we might want to account for tail behaviour, we might consider restricting the range of approximation to a bounded linear subspace of $\mathbb{R}^d$. This can be done by introducing a linear transformation of the form

$$x = b + Cy, \qquad y \in [-1, 1]^d \tag{12}$$

with Jacobian $\mathcal{J}_L = |C|$. For example, if we use the diagonal transformation

$$x_i = \frac{1}{2}[(b_i + c_i) + y_i(b_i - c_i)], \qquad b_i > c_i \tag{13}$$

with Jacobian $\mathcal{J}_L = \prod_{i=1}^{d} \frac{1}{2}(b_i - c_i)$, then $x_i \in [c_i, b_i]$. More generally, by using a non-diagonal transformation, we can take advantage of tilted axes or asymmetries in $\varphi(x)$.

Selection of an n-point Legendre quadrature generates $N = n^d$ product nodes and weights $\{(y_i^L, w_i^L)\}_{i=1}^N$ that are transformed into $\{(x_i, w_i)\}_{i=1}^N$ by Equation (12), together with $w_i = \mathcal{J}_L w_i^L \varphi(x_i)$. It follows that the distance measure $f_J(a_J)$ in Equation (8) is approximated by

$$\hat{f}_J(a_J) = \frac{1}{2} \sum_{i=1}^{N} w_i \left[\ln \varphi(x_i) - \ln k_J(x_i, a_J)\right]^2 . \tag{14}$$

Minimization of $\hat{f}_J(a_J)$ with respect to a_J is discussed in section 2.5 below. One potentially important computational advantage of Legendre quadratures as well as Hermite quadratures discussed next, is that the nodes and weights $\{(x_i, w_i)\}_{i=1}^N$ remain unchanged across all J's. This is not the case with Importance Sampling in Equation (10), or with Hermite mixture quadratures in section 2.3.3 below.

2.3.2 Hermite

The use of Hermite quadratures offers the advantage that it operates on $\mathbb{R}^d$ though it requires attention since it relies on a Gaussian thin tail weight function. It is particularly attractive when $\varphi(x)$ itself includs a Gaussian kernel, say

$$\varphi(x) = \phi(x)F(x) \tag{15}$$

with

$$\phi(x) = \exp\left[-\frac{1}{2}(x - m_0)' H_0 (x - m_0)\right] \tag{16}$$

and $F(x)$ typically well-behaved. In such a case we can rely on a transformation of the form

$$x = m_0 + \sqrt{2}P_0 y, \qquad \text{with} \quad P_0^{'} H_0 P_0 = I_d \tag{17}$$

and Jacobian $\mathcal{J}_H = 2^{d/2}|P_0|$. $\phi(x)$ is then transformed into the Hermite weight function $\exp(-y^{'}y)$. The Hermite nodes and weights $\{(y_i^H, w_i^H)\}_{i=1}^N$ are transformed into $\{(x_i, w_i)\}_{i=1}^N$ by Equation (17) together with $w_i = \mathcal{J}_H w_i^H$ and $\hat{f}_J(a_J)$ is estimated according to the Equation (14).

Actually, we can use Hermite even when $\varphi(x)$ does not include a Gaussian kernel provided we pay attention to tail behaviour. Specifically, by introducing an auxiliary kernel $\phi(x)$ of the form given by Equation (16) we can rewrite $f_J(a_J)$ as

$$f_J(a_J) = \frac{1}{2}\int \left[\ln \varphi(x) - \ln k_J(x, a_J)\right]^2 \left[\frac{\varphi(x)}{\phi(x)}\right] \phi(x)dx. \tag{18}$$

This equation is then evaluated using the Equation (14) with the following adjustments: we now use Hermite nodes and weights and the corresponding adjusted weights w_i are given by

$$w_i = \mathcal{J}_H w_i^H \left[\frac{\varphi(x_i)}{\phi(x_i)}\right]. \tag{19}$$

It is then critical that the ratios $\varphi(x_i)/\phi(x_i)$ remain sufficiently well-behaved (at minimum for all x_i's). Laplace approximations are often used to construct Gaussian kernel approximations. However, they can produce tails that are too thin and induce unacceptably large variations in the weights w_i. We recommend instead using moment approximations for m_0 and H_0, following a procedure presented in section 3 to compute initial values.

2.3.3 Mixture-Hermite

A computationally more intensive but potentially more accurate procedure consists of using a J-term mixture approximation as weight function in step J. Specifically, $f_J(a_J)$ is rewritten as

$$f_J(a_J) = \frac{1}{2}\sum_{j=1}^{J} e^{\delta_j^o} \int \left[\ln \varphi(x) - \ln k_J(x, a_J)\right]^2 \nu(x, a_J^o) k_J(x, \alpha_J^o)dx. \tag{20}$$

with

$$\nu(x, a_J^o) = \frac{\varphi(x)}{k_J(x, a_J^o)}, \qquad j : 1 \to J \tag{21}$$

where $a_J^o = \{\alpha_j^o, \delta_j^o\}_{j=1}^J$, are set (and kept fixed) at the initial values selected for the a_J optimization. Indeed, we do not recommend using an EIS type fixed-point optimization sequence for $\hat{a}_J$ since, in particular, the optimal mixture that obtains at step J will be replaced by a new one at step $J+1$ (as long

as we keep increasing J). An obvious choice for $a_J^o = \{\alpha_j^o, \delta_j^o\}_{j=1}^{J-1}$ for $J > 1$ consists of the optimal $\hat{a}_{J-1}$ obtained at step $J-1$, while for (α_J^o, δ_J^o) we can use the initial values for step J obtained as described in section 2.6.2 below. Actually, for $J > 1$, we can run the summation in Equation (20) from $j = 1$ to $J-1$, ignoring the new term. Both alternatives are covered by Equation (20) if we run summation from $j = 1$ to J_M, where $J_M = 1$ for $J = 1$ and either J or $J-1$ for $J > 1$.

Next, we apply the transformation in Equation (17) indexed by j to each term in the summation. This produces a new set of nodes and weights that are given by

$$x_{ij} = m_0^j + \sqrt{2} P_0^j y_i \tag{22}$$

$$w_{ij} = e^{\delta_j^0} w_j \nu(x_{ij}, a_J^o) \tag{23}$$

for $i : 1 \to N$ and $j : 1 \to J_M$. The estimate of $f_J(a_J)$ is then given by

$$\hat{f}_J(a_J) = \frac{1}{2} \sum_{j=1}^{J_M} \sum_{i=1}^{N} w_{ij} \left[\ln \varphi(x_{ij}) - \ln k_J(x_{ij}, a_J) \right]^2 . \tag{24}$$

Potential advantages of that procedure are twofold. As J increases, $k_J(x, a_J)$ provides a closer approximation to $\varphi(x)$ so that the variance of the ratios $\nu(x, a_J^o)$ is expected to decrease significantly thereby alleviating the thin tail problem inherent to Hermite. Also the number of nodes is now given by NJ_M and is, therefore, proportional to the number of auxiliary parameters in a_J. Thus it is possible to reduce the number N of grid points accordingly. A significant drawback is that each J iteration relies upon a new grid, in sharp contrast with the Legendre and Hermite when the grid remains the same for all J's.

2.4 Identification

In it well known that Maximum Likelihood (thereafter ML) estimation of mixtures raises important issues of identifiability and regularity. See Titterington et al (1985, section 3.1) or Frühwirth-Schnatter (2006, section 1.3). These are three main issues: (i) mixtures are invariant relative to a permutation (relabeling) of their components; (ii) parameters of a component with (near) zero probability or of two equal components are not (or poorly) identified - this is referred to as "overfitting"; and (iii) determination of the number of components is complicated by the fact that standard asymptotic theory does not apply when parameters lie at the boundary of the parameter space. See McLachlan and Peel (2000, section 6.1) or Kasahara and Shimotsu (2015).

Relabeling or permutation appear to have no practical implications for our algorithm. While it certainly can happen, it is inconsequential for our gradient minimization of $f_J(a_J)$. We have never faced a convergence problem that could be attributed to relabeling. Initially, we did incorporate in our algorithm an

ordering of the means but found out that is complicates programming and does not affect or even accelerate convergence. Failure of regularity conditions is irrelevant in a framework where we discuss approximating a known density kernel and when, as we discuss next, addition of new terms is linked to further reductions in the distance measure $f_J(a_J)$.

Overfitting is obviously an issue but one that is actually easy to address. As discussed in the Appendix, gradients are proportional to the mixture weights $e^{\delta_j^o}$ to the extent that optimization will inevitably be problematic for any new term with a (relatively) very low weight. However, such terms would minimally contribute to lowering further $f_J(a_J)$. Thus, as discussed next, low weight is one of the stopping criterion that can be implemented.

2.5 Minimization of the distance measure

In order to minimize the distance measure $f_J(a_J)$ in Equation (8), more specifically its quadrature estimates in Equation (14), (18) or (24), we can take advantage of the fact that the first and second order derivatives of $\ln k_J(x, a_J)$ with respect to a_J obtain analytically. Thus, we can use numerical optimizers that rely upon analytic gradients and, possibly, Hessians. After extensive initial experimentation, we found out that a quasi-Newton method using analytic gradient is numerically efficient for minimizing $f_J(a_J)$. The expressions for the analytic gradient of $f_J(a_J)$ are derived in Appendix.

In addition to supplying subroutines to analytically evaluate $f_J(a_J)$ and its gradient, we also need to provide initial values and a diagonal scaling matrix. Initial values are derived in the next section. As for scaling, we found that the default option (all diagonal entries set to 1) works perfectly fine as long as $\varphi(x)$ is approximately normalized in order to avoid large imbalances with $k_J(x, a_J)$. While such normalization was not needed for the examples presented below, an obvious solution consists of dividing $\varphi(x)$ by G_0, an initial quadrature estimate of its integral as presented next.

2.6 Initial values

Numerical minimization of $f_J(a_J)$ in step J requires initial values for $a_J = \{\mu_j, R_j, \delta_j\}_{j=1}^J$ in Equation (3). Thus, for $J = 1$, we need to provide initial $(\mu_1^o, R_1^o, \delta_1^o)$. For $J > 1$, it is natural to define the new initial value of a_J^o as $a_J^o = \hat{a}_{J-1} \cup (\mu_J^o, R_J^o, \delta_J^o)$, where $\hat{a}_{J-1}$ denotes the optimal mixture parameters obtained at step $J - 1$ (with a minor proportional adjustment to the mixture weight).

A fairly common practice in the literature surveyed in Introduction, consists of relying upon (local) Laplace approximations to construct μ_J^o and $H_J^o = R_J^o R_J^{o'}$. For example, Ardia et al (2009) define μ_J^o as the (global) maximum of the importance sampling log ratio

$$\ln \nu_{J-1}(x, \hat{a}_{J-1}) = \ln \varphi(x) - \ln k_{J-1}(x, \hat{a}_{J-1}), \tag{25}$$

and use minus its Hessian for H_j^o. Bornkamp (2011) applies the same idea to the log difference $\ln r_{J-1}(x)$, with

$$r_{J-1}(x) = \varphi(x) - k_{J-1}(x, a_{J-1}), \tag{26}$$

where $r_{J-1}(x)$ has to be bounded below by some $\epsilon > 0$ to avoid problems computing its logarithm. We experimented with Bornkamp's method and found out that it works overall quite well.

However, we now rely on a different approach to construct initial values that takes advantage of the fact that Gaussian quadratures can be used to compute moments (whether truncated or not) directly. The advantage of this procedure is twofold: (i) it replaces local Laplace approximations by global ones, a concept that is central to the EIS principle introduced by Richard and Zhang (2007); and (ii) it relies exclusively upon function evaluations that were already produced using the step $J-1$ Gaussian grid, while Laplace approximations require new function evaluations for the mode and Hessian. Thus, the computation of initial values relies upon integrals of the form:

$$H = \int_D h(x)\varphi(x)dx. \tag{27}$$

Under Legendre and Hermite rules, the computation of H relies upon the fixed grid $(x_i, w_i)_{i=1}^N$ associated with the selected rule. Under the mixture approach for $J > 1$, the grid consists of the grids associated with the $J-1$ individual Gaussian kernels in $k_{J-1}(x, \hat{a}_{J-1})$. For the ease of notation, we run the summation over i from 1 to M, where M is either N (Legendre, Hermite) or $(J-1)N$ (mixture for $J > 1$). Let $\nu(x)$ denote the ratio between $\varphi(x)$ and the selected weight function. It is given by

$$\text{Legendre}: \quad \nu(x) = 1 \tag{28a}$$

$$\text{Hermite}: \quad \nu(x) = \varphi(x)/\phi(x), \quad \text{with} \quad \phi(x) \quad \text{defined in (18)} \tag{28b}$$

$$\text{Mixture}(J > 1): \quad \nu(x) = \varphi(x)/k_{J-1}(x, \hat{a}_{J-1}) \tag{28c}$$

The quadrature estimate of H is then given by

$$\hat{H}_N = \sum_{i=1}^{M} \tilde{w}_i h(x_i), \tag{29}$$

where $\tilde{w}_i$ denotes the adjusted weight

$$\tilde{w}_i = w_i \nu(x_i). \tag{30}$$

Next, we describe how formulas (28)-(30) are used to construct the initial values $a_J^o = \{\mu_j^o, R_j^o, \delta_j^o\}_{j=1}^J$.

2.6.1 Initial values for step $J = 1$

Under Legendre and Hermite rules, we compute initial values for (μ_1^o, R_1^o) as follows:

$$\mu_1^o = \sum_{i=1}^{M} w_i^* x_i \tag{31}$$

$$\Sigma_1^o = \sum_{i=1}^{M} w_i^* (x_i - \mu_1^o)(x_i - \mu_1^o)' \tag{32}$$

with

$$w_i^* = \frac{\tilde{w}_i}{\sum_{j=1}^{M} \tilde{w}_j}. \tag{33}$$

and R_1^o obtaining from the Cholesky factorization of $H_1^o = \Sigma_1^{o-1} = R_1^o R_1^{o'}$.

As for δ_1^o, we equate the initial estimate of G_0 with $(2\pi)^{d/2}$, the integrating factor of $k(x, \alpha_1^o)$. Thus

$$\delta_1^o = \ln\left(\sum_{i=1}^{M} \tilde{w}_i\right) - \frac{d}{2}\ln 2\pi. \tag{34}$$

For the mixture approach, we use either Legendre or Hermite, as described above, to produce the initial step $J = 1$ mixture.

2.6.2 Initial values for step $J > 1$

As already mentioned, the initial values for step $J > 1$ essentially consist of the optimal $\hat{a}_{J-1}$ obtained at step $J - 1$ complemented by initial values for the added term:

$$a_J^o \simeq \hat{a}_{J-1} \cup (\mu_J^o, R_J^o, \delta_J^o) \tag{35}$$

with a downward adjustment for $(\hat{\delta}_j)_{j=1}^J$. The latter is justified by the fact that the integrating factor of the successive mixture $k_J(x, \hat{a}_J)$ all approximate the same (unknown) constant G. Thus the addition of a new term with $\exp(\delta_J^o) > 0$ should result in a reduction of the current $\hat{\delta}_j$'s. We experimented with a variety of rules of thumb to select δ_J^o. Based on the observation that new terms generally exhibit decreasing $\hat{\delta}_j$'s, we adopted the following simple rule that works consistently well:

(i) Define δ_J^* as the smallest of the current $\hat{\delta}_j$'s:

$$\delta_J^* = \min \hat{\delta}_j, \quad \text{for} \quad j = 1, ..., J-1 \tag{36}$$

(ii) Compute an adjustment ratio $\theta_J < 1$ defined as

$$\theta_J = \left(\sum_{j=1}^{J} e^{\hat{\delta}_j}\right)\left(e^{\delta_J^*} + \sum_{j=1}^{J-1} e^{\hat{\delta}_j}\right)^{-1} \tag{37}$$

(iii) The step J initial weights are then given by

$$\begin{aligned} \delta_j^o &= \hat{\delta}_j + \ln \theta_J \quad \text{for} \quad j = 1, ..., J-1 \\ \delta_J^o &= \delta_J^* + \ln \theta_J \end{aligned} \tag{38}$$

Given θ_J, we define the truncated density kernel

$$\begin{aligned} \kappa_{J-1}(x) &= \varphi(x) - \theta_J k_{J-1}(x, \hat{a}_{J-1}) \quad \text{if positive} \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{39}$$

and the initial values for (μ_J^o, R_J^o) obtain as for step 1, with $\varphi(x)$ replaced by $\kappa_{J-1}(x)$. Even with $\theta_J < 1$, there remain a theoretical possibility that $\kappa_{J-1}(x)$ could have a sharp peak (relative to the quadrature grid) to the effect that the (non-negative) Σ_J^o could be (near) singular. We have not yet encountered that eventuality but it would be trivial to fix either by adding to Σ_J^o a small positive scalar multiple of the identity matrix I_d, or by reverting to a Laplace approximation of $\ln \kappa_{J-1}(x)$, where $\kappa_{J-1}(x)$ would then be bounded below by $\epsilon > 0$, as in Bornkamp (2011).

3 Test cases

In this section we present three test cases taken from the literature and highlighting key features of our approach. The first is taken from Gilks et al (1998) (also used in Bornkamp, 2011) with a bivariate target mixture and illustrates the importance of full re-optimization of the approximating mixture with the introduction of each new term. The second case is taken from Azzalini and Dalla Valle (1996). The target is bivariate skew-distribution representing a class of densities of growing importance in econometrics. It also illustrates the importance of reducing the dimension of the kernel that has to be approximated as mixtures do suffer from an obvious curse of dimensionality, to be discussed further below. The last case discusses a mixture approximation to the density of a $\log \chi_1^2$ variable. As we discuss in section 4, such approximations provide an important tool to construct a mixture filtering approach to stochastic volatility models.

3.1 Mixture of three bivariate normal distributions

Example 2 in Bornkamp (2011) applies the iterated Laplace algorithm to the following bivariate target mixture, originally used in Gilks et al (1998):

$$\varphi(x) = \sum_{i=1}^{3} \pi_i f_N(x|\mu_i, \Sigma_i), \tag{40}$$

Table 1 Initial and terminal values for approximating the mixture of three bivariate normal distributions

J	j	initial values			terminal values		
		$\exp(\delta_i^o)$	μ_i^o	Σ_i^o	$\exp(\hat{\delta}_i)$	$\hat{\mu}_i$	$\hat{\Sigma}_i$
1	1	0.027	$\begin{pmatrix}-0.336\\-0.336\end{pmatrix}$	$\begin{pmatrix}5.155 & 4.159\\4.159 & 5.155\end{pmatrix}$	0.204	$\begin{pmatrix}-0.298\\-0.298\end{pmatrix}$	$\begin{pmatrix}6.110 & 4.936\\4.936 & 6.110\end{pmatrix}$
		$f_1(a_1^o) = 59.131$			$f_1(\hat{a}_1) = 18.381$		
2	1	0.102	$\begin{pmatrix}-0.298\\-0.298\end{pmatrix}$	$\begin{pmatrix}6.110 & 4.936\\4.936 & 6.110\end{pmatrix}$	0.757	$\begin{pmatrix}1.447\\1.447\end{pmatrix}$	$\begin{pmatrix}2.365 & 0.610\\0.610 & 2.365\end{pmatrix}$
	2	0.102	$\begin{pmatrix}-0.332\\-0.333\end{pmatrix}$	$\begin{pmatrix}5.215 & 4.220\\4.220 & 5.215\end{pmatrix}$	0.399	$\begin{pmatrix}-2.671\\-2.671\end{pmatrix}$	$\begin{pmatrix}1.751 & 1.640\\1.640 & 1.751\end{pmatrix}$
		$f_2(a_2^o) = 18.431$			$f_2(\hat{a}_2) = 0.967$		
3	1	0.562	$\begin{pmatrix}1.447\\1.447\end{pmatrix}$	$\begin{pmatrix}2.365 & 0.610\\0.610 & 2.365\end{pmatrix}$	0.330	$\begin{pmatrix}2.000\\2.000\end{pmatrix}$	$\begin{pmatrix}1.000 & -0.900\\-0.900 & 1.000\end{pmatrix}$
	2	0.296	$\begin{pmatrix}-2.671\\-2.671\end{pmatrix}$	$\begin{pmatrix}1.751 & 1.640\\1.640 & 1.751\end{pmatrix}$	0.330	$\begin{pmatrix}-3.000\\-3.000\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.900\\0.900 & 1.000\end{pmatrix}$
	3	0.296	$\begin{pmatrix}-0.059\\-0.059\end{pmatrix}$	$\begin{pmatrix}5.160 & 4.066\\4.066 & 5.160\end{pmatrix}$	0.340	$\begin{pmatrix}0.000\\0.000\end{pmatrix}$	$\begin{pmatrix}1.000 & 0.000\\0.000 & 1.000\end{pmatrix}$
		$f_3(a_3^o) = 1.076$			$f_3(\hat{a}_3) = 2.136E-8$		

with $(\pi_1, \pi_2, \pi_3) = (0.34, 0.33, 0.34)$, $\mu_1^{'} = (0, 0)$, $\mu_2^{'} = (-3, 3)$, $\mu_3^{'} = (2, 2)$, $\Sigma_1 = \left(\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right)$, $\Sigma_2 = \left(\begin{smallmatrix}1 & 0.9\\0.9 & 1\end{smallmatrix}\right)$, $\Sigma_3 = \left(\begin{smallmatrix}1 & -0.9\\-0.9 & 1\end{smallmatrix}\right)$. Bornkamp's algorithm constructs the mixture approximation sequentially as we do but does not re-optimize their Laplace moments. Thus it cannot replicate the target. Actually, it ends producing a five-term mixture approximation whose means and standard deviations are all within less than 1% of those of the moments of the target density. In sharp contrast, our algorithm reproduces exactly the target density (up to the optimizer's stopping rule). In order to illustrate how it works, we reproduce in Table 1 initial and final values for the three successive iterations using Legendre rule on the range $[-6, 6]^2$, though any reasonable range will deliver the same perfect fit. Similar results obtain under the Hermite and mixture approach.

3.2 Skew-Normal density

Multivariate skew-distributions are gaining importance in stochastic frontier analysis and sample selection models. Stochastic frontier analysis models have been using skewness as an intrinsic characteristic to measure technical inefficiency. The skewed shape of the error term in the stochastic frontier problem arises from its composite structure, consisting of two separate error components – a symmetric measurement error and an inefficiency factor, defined to be one-sided. Dealing with production frontiers corresponding to firms producing multiple outputs ultimately led to the system of stochastic frontier equations, where the multivariate skewed distribution is applied to model the

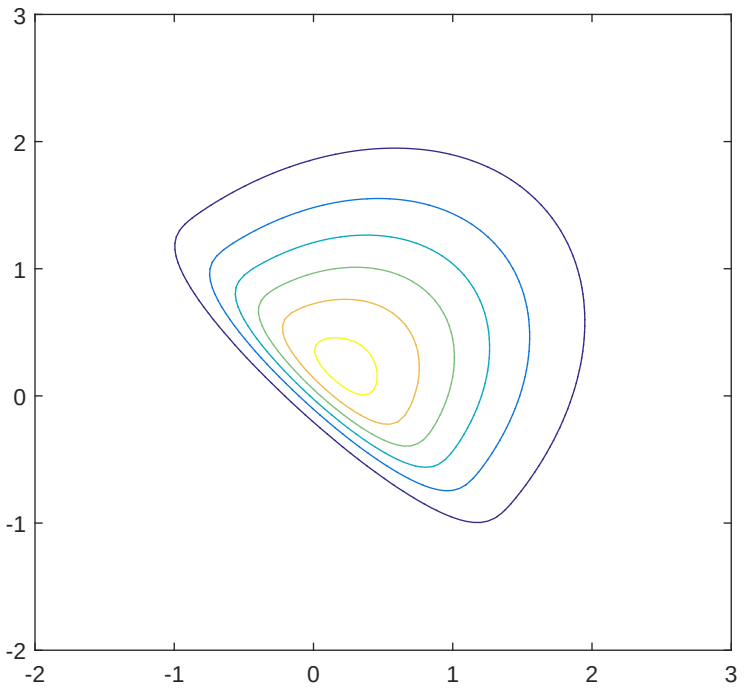

Fig. 1 Contour plot SN_2 for $\omega = 0.3$ and $\delta = 0.8$.

composite error (Domínguez-Molina et al, 2007; Ferreira and Steel, 2007). In sample selection models, skew-distributions are used to mitigate the effects of distributional misspecifications. The distribution of many economic outcomes (e.g., wages) is likely to be skew in the population, before selection. Further, the skewness in outcomes could be induced by the selection itself as a hidden truncation. These considerations led to developing parametric sample selection models for skew outcomes such as in Marchenko and Genton (2012) and Ogundimu and Hutton (2016).

Our second test case is related to dealing with multivariate skew-distributions and consists of the following bivariate skew-normal density taken from Azzalini and Dalla Valle (1996):

$$\varphi(x) = \frac{1}{\pi}\left[|\Omega|^{-\frac{1}{2}} \exp(-\frac{1}{2}x'\Omega^{-1}x)\right] \Phi(\alpha\iota' x), \tag{41}$$

where Φ denotes the standardized Normal cdf, $\iota' = (1,1)$, $\Omega = \left(\begin{smallmatrix} 1 & \omega \\ \omega & 1 \end{smallmatrix}\right)$, and $\alpha = \delta(1-\omega)\{(1-\omega^2)[1-\omega^2-2\delta^2(1-\omega)]\}^{-\frac{1}{2}}$, with $\omega = 0.3$ and $\delta = 0.8$. Its skewed contour plot is presented in Figure 1.

Table 2 Mixture moments' comparison for approximation of the bivariate skew-normal density

	Hermite (5-term)		Legendre	Mixture-Hermite
	Orthogonal	Cholesky	(5-term)	(7-term)
μ	$\begin{pmatrix}0.6399\\0.6399\end{pmatrix}$	$\begin{pmatrix}0.6368\\0.6383\end{pmatrix}$	$\begin{pmatrix}0.6392\\0.6351\end{pmatrix}$	$\begin{pmatrix}0.6377\\0.6380\end{pmatrix}$
Σ	$\begin{pmatrix}0.5904 & -0.1096\\-0.1096 & 0.5904\end{pmatrix}$	$\begin{pmatrix}0.5922 & -0.1042\\-0.1042 & 0.5904\end{pmatrix}$	$\begin{pmatrix}0.5922 & -0.1094\\-0.1094 & 0.5942\end{pmatrix}$	$\begin{pmatrix}0.5916 & -0.1086\\-0.1086 & 0.5912\end{pmatrix}$
Nodes	90 univar.nodes	28×28^a	28×28^a	$J10\times10^a$
Comp. time[b]	0.45	2.54	2.54	2.84

[a] product rule.
[b] seconds.

Since $\varphi(x)$ already includes a Gaussian kernel, it is natural to apply Hermite rule. The two obvious Ω factorizations leading to transformation (17) are the Cholesky and orthogonal ones. The corresponding transformations are given by

$$x = \sqrt{2}\begin{pmatrix}1.00 & 0.00\\0.30 & \sqrt{0.91}\end{pmatrix} y, \qquad (42a) \qquad x = \sqrt{2}\begin{pmatrix}\sqrt{0.65} & \sqrt{0.35}\\\sqrt{0.65} & -\sqrt{0.35}\end{pmatrix} y, \qquad (42b)$$

respectively. It turns out that the orthogonal transformation produces a much simpler expression for the transformed target that is given by

$$\varphi(y) = \frac{2}{\pi}\varphi_1(y_1)\varphi_2(y_2), \tag{43}$$

with $\varphi_1(y_1) = \Phi(8\sqrt{2}y_1)\exp(-y_1^2)$, and $\varphi_2(y_2) = \exp(-y_2^2)$. Therefore, we only need to construct a univariate mixture approximation $k_1(y_1, \hat{a})$ for $\varphi_1(y_1)$ and the corresponding bivariate mixture approximation for $\varphi(y)$ obtains as

$$k(y, \hat{a}) = k_1(y_1, \hat{a})\varphi_2(y_2), \tag{44}$$

to be transformed back into a mixture approximation for $\varphi(y)$ by the inverse transformation (42b). We can also apply Hermite quadrature to compute the "true" moments of y_1 and, therefore, those of x. Using 1,000 quadrature points since $\Phi(8\sqrt{2}y_1)$ is very tight, we find that $\mu_1 = \mu_2 = 0.63830765$, $\sigma_{11} = \sigma_{22} = 0.59256335$ and $\sigma_{12} = \sigma_{21} = -0.10743665$. Both transformations in (42) produce 5-term mixture approximations with plot contours that are virtually indistinguishable from that of $\varphi(y)$ in Figure 1. The corresponding mixture moments under both transformations are given in Table 2. The orthogonal transformation produces fairly accurate results as expected, though it requires additional algebraic transformations. It illustrates the importance of exploring dimension-reducing transformations both for accuracy and to reduce the curse of dimensionality inherent to finite mixtures. For comparison, we apply Legendre and Mixture-Hermite rules to the test case of bivariate skew-normal density and obtain the comparable 5-term mixture approximation for Legendre rule and 7-term mixture approximation for Mixture-Hermite rule with the corresponding mixture moments, presented in Table 2.

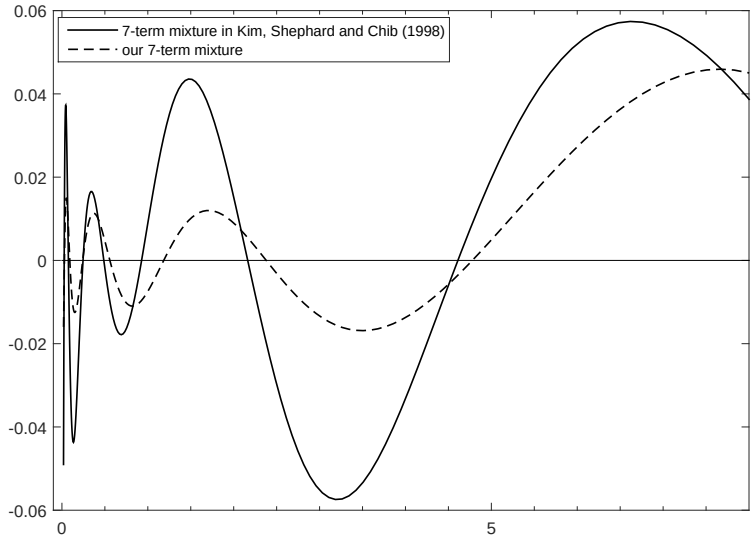

Fig. 2 The log of the ratio of the χ^2_1 density to the mixture approximation.

3.3 Basic stochastic volatility model

A density kernel for a $\log \chi^2_1$ random variable is given by

$$\varphi(x) = \exp\left[\frac{1}{2}(x - e^x)\right], \tag{45}$$

As is well known and discussed further in section 4 below, this kernel plays a central role in likelihood (filtering) evaluations of a number of Stochastic Volatility (thereafter SV) models. Since $\varphi(x)$ is significantly skewed, it is natural to consider approximating it by a finite Gaussian mixture. One such mixture is proposed by Kim et al (1998, Equation (10) and Table 4) and is obtained by "using a non-linear least squares program to move the weights, means and variances around until the answers were satisfactory". Adjusting for their mean shift of 1.2704, we use their parameter values as initial values for a direct 200 point Legendre minimization of $\hat{f}_7(a_7)$ in Equation (14) over the range [-20, 4]. The comparable results are reported in Table 2 and Figure 2.

Optimization has reduced the distance measure f_7 by a factor 19. Since $f_J(a_J)$ is (approximately) proportional to the Importance Sampling variance of the corresponding IS ratios, such large reductions would result in equally large reductions in the number of draws in IS applications.

Table 3 Mixture approximation of the $\log \chi_1^2$ kernel

	initial values			optimal values		
	π_i	μ_i	σ_i^2	π_i	μ_i	σ_i^2
1	0.00730	-10.12999	5.795960	0.01661	-6.44535	13.58034
2	0.00002	-8.56686	5.179500	0.00002	-8.58075	3.70735
3	0.10556	-3.97281	2.613690	0.08720	-3.59047	4.86088
4	0.25750	-1.08819	1.262610	0.20824	-1.38055	2.09610
5	0.34001	0.61942	0.640090	0.30992	0.23027	0.99504
6	0.24566	1.79518	0.34023	0.27751	1.43183	0.51327
7	0.04395	2.77786	0.16735	0.10073	2.37341	0.28337
	$f_7(a_7^o)$=6.8544E-003			$f_7(\hat{a}_7)$=3.6942E-004		

4 Future research

Our generic procedure to construct finite Gaussian mixture approximations to analytically intractable density kernels provides alternative numerical solutions to a wide range of problems in statistics, economics and finance. We outline below three ongoing projects for which we have already produced promising initial results. We also discuss extensions to non-Gaussian mixtures.

4.1 Filtering

Dynamic state space models are increasingly widely used in sciences, including economics. When the latent state and the measurement process are both linear Gaussian, the Kalman Filter provides operational fully analytical solutions. When this is not the case, Particle Filters (hereafter PF's) that rely upon Sequential Important Sampling and extensions thereof are commonly used to produce approximations to the relevant densities (filtering, predictive and likelihood) in the form of discrete mixtures of Dirac measures (referred to as swarms of particles). PF's are widely applicable but also suffer from potential problems, foremost degeneracy and sample impoverishment (see e.g. Ristic et al, 2004, for an in-depth presentation of particle filters with emphasis on tracking applications). Various extensions of the baseline PF algorithm have been produced to enhance its numerical efficiency (see e.g. Pitt and Shephard, 1999, the collection of papers in Doucet et al, 2001; see also section II.D in Cappé et al, 2007 for advances in Sequential Monte Carlo, of which the Mixture Kalman filter is directly relevant to the present project). It applies to a broad range of state space models that consist of a linear Gaussian latent state process combined with a non-linear or non-Gaussian measurement process. It combines Kalman filtering for the state part, and particle filtering for the measurement part. Our ongoing project consists of replacing the latter by a Gaussian mixture approximation of the measurement density. Doing so essentially amounts to constructing a mixture extension of the Kalman filter.

In a nutshell, it operates as follows. The non-linear or non-Gaussian measurement densities are approximated by finite Gaussian mixtures. In period t, one inherits a period $t-1$ filtering mixture approximation, which is combined with the state linear Gaussian transition in order to produce a predictive mixture approximation. The latter is then multiplied by the measurement mixture approximation. Assuming we are relying upon J-term mixtures, this product takes the form of a J^2-term mixture that can in turn be approximated by a J-term mixture (by selecting the J terms with highest probability, re-scaling them into initial values and re-optimizing). The likelihood then obtains as the analytical integrating constant of the mixture kernel and the period t filtering density as the normalized version of that same mixture. Moreover, once we have run the forward filtering algorithm, it is possible to run it backward in order to produce smooth (mixture bound) estimates of the state variables.

Unsurprisingly, there is a fair amount of analytical details to be cleaned up in order to produce a generic mixture extension of the Kalman filter but we have already tested it on a univariate baseline stochastic volatility application taken from Liesenfeld and Richard (2006). That application offers the critical advantage that the period t measurement density obtains as a linear transformation of a canonical $\log \chi_1^2$ density, whose mixture approximation was presented in section 3.3 and needs to be computed only once. The application consists of a sample of 945 weekly exchange rates for the British pound against the US dollar. Using mixture approximations, we obtained the following values for the log-likelihood at the ML parameter values: -918.62 (7-term mixtures) and -918.61 (8-term mixtures). For comparison, Liesenfeld and Richard (2006, Table 1, column 2) report an EIS estimate of -918.60. Moreover, 100 MC-EIS replications produce a mean of -918.66 with a standard deviation of 0.026 and a range (-918.72, -918.59). Obviously, our mixture estimates are non-stochastic but their high numerical accuracy is illustrated by the near identical values obtained under 7- and 8-term mixtures.

The results of that pilot application are extremely encouraging and we are currently developing a generic multivariate mixture extension of the baseline Kalman filter (log)-likelihood estimation as well as filtered and smooth state estimates. Our plan is to test this mixture algorithm to the three-dimensional state-space stochastic volatility model for inflation, as analyzed by Stock and Watson (2007, section 3).

4.2 Mixture approximations of non-parametric density estimates

Finite Gaussian mixtures are used increasingly as approximations for nonparametric kernels (see, e.g., Scott and Szewczyk, 2001). The papers by Han et al (2008) and Wang and Wang (2015) include useful surveys of the recent literature to that effect as well as new proposals for large reductions in the number of components. The most commonly proposed method consists of sequential reductions of the number of terms based upon a variety of clustering procedures. We propose instead to apply our algorithm directly to the nonparametric ker-

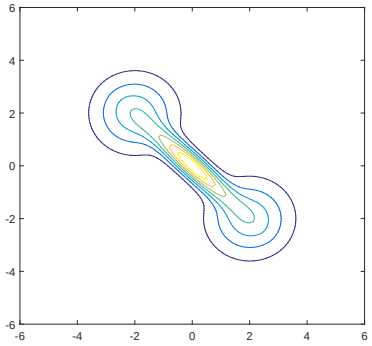

Fig. 3 "Dumbbell" density

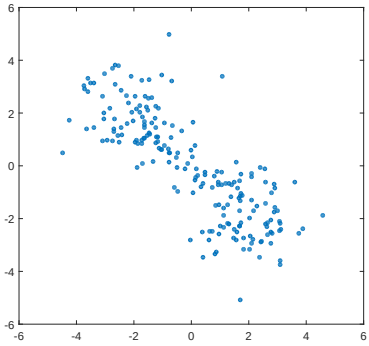

Fig. 4 Data points drawn from "dumbbell" density

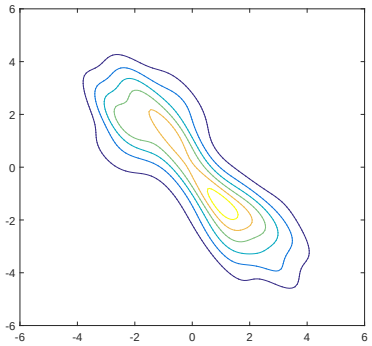

Fig. 5 Duong's (2007) kernel density estimates for "dumbbell" data

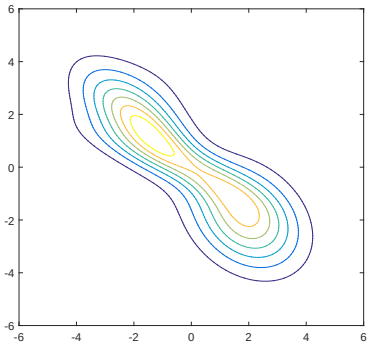

Fig. 6 6-term mixture approximation

nel as target, adding terms one by one using our distance measure to assess the goodness of fit of the mixture approximation. As a pilot illustration of the potential of such procedure, we used a simple example taken from Duong (2007), where the author constructs nonparametric density kernels for a data set consisting 200 i.i.d. draws from a "dumbbell" (unimodal) density given by the normal mixture

$$\frac{4}{11}\left[N\left(\begin{pmatrix}-2\\2\end{pmatrix},I_2\right)+N\left(\begin{pmatrix}2\\-2\end{pmatrix},I_2\right)\right]+\frac{3}{11}N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}0.80 & -0.72\\-0.72 & 0.80\end{pmatrix}\right).$$

The density is illustrated in Figure 3, whereas the 200 data points drawn from this density are plotted in Figure 4.

We applied our algorithm to produce a 6-term mixture approximation to Duong's (2007) plug-in nonparametric kernel estimate. The contours for the Duong's (2007) nonparametric estimate are presented in Figure 5, whereas Figure 6 illustrates our 6-term mixture approximation. Here again, the results of this pilot application are very promising. Our current objective is that of

producing an algorithm applicable to large data sets, where dramatic reductions in the number of terms and clustering will be critical for analysis. We aim at achieving high numerical efficiency for such simplification exercises. A critical step toward that objective consists of replacing the quadrature grid by the data, reinterpreted as equal weight draws from the nonparametric kernel estimate to be approximated. Initial value calculations are to be adjusted accordingly.

4.3 Sparse grids

The product rules used for the numerical evaluation of the distance measure in Equation (8) suffer from an obvious "curse of dimensionality". As explained by Heiss and Winschel (2008, section 2.4), the exponential growth of computational costs as a function of the dimension d originates from the fact that the product rule is exact for a tensor product of univariate polynomials, not for polynomials of bounded total order. The concept of sparse grids combines univariate rules in such a way that it is exact for complete polynomials of a given order with computational costs rising considerably slower than exponentially. The basic idea originates from Smolyak (1963) providing a generic procedure for multivariate generalizations of univariate operators (see Bungartz and Griebel, 2004 for a detailed presentation and Heiss and Winschel, 2008 for a self-contained description of how to construct sparse grids).

We have started exploring how to produce a sparse grid version of our mixture algorithm. An immediate problem arises from the fact that a significant percentage (typically close to 50%) of the weights associated with the nodes are negative. It follows that the baseline distance measure in Equation (23) is no longer bounded below by zero and, consequently, that its minimization generally fails. An obvious remedy consists of replacing the negative weights in (23) by their absolute values. This produces an objective function that can no longer be interpreted as an approximation of the sampling variance of the IS ratios in Equation (7) but one that can still be interpreted as a distance measure.

Our next step will be that of adjusting our procedure to compute initial values. While using Laplace approximations remains possible, it can be computationally inefficient, especially as the dimension d gets larger and sparse grid points increasingly dispersed. Our truncated moments approach avoids additional target evaluations outside of the grid but negative weights remain problematic as they could occasionally produce non-positive truncated initial covariance matrices.

For illustration purposes, we rerun the bivariate skew-normal density example presented in section 3.2 under sparse grids with Laplace initial values. We obtain the following results with the Cholesky transformation and 200 sparse-grid nodes:

$$\mu = \begin{pmatrix} 0.63657460 \\ 0.63658100 \end{pmatrix},\ \Sigma = \begin{pmatrix} 0.58987717 & -0.10914464 \\ -0.10914464 & 0.58987327 \end{pmatrix},$$

which are similar to those reported in section 3.2. Computing time is 0.53 seconds. Thus, the use of sparse grids provides a very promising lead for extending our algorithm beyond dimension two or three.

4.4 Other mixture types

While Gaussian mixtures are by far the most commonly used, other types are worthy of consideration. For example, Hoogerheide et al (2007) and Hoogerheide et al (2012) use mixtures of Student-t kernels with one degree of freedom to approximate targets with fat tails. Titterington et al (1985, Table 2.1.3, pages 6-21) provide an extensive list of applications, many with non-Gaussian mixture types (von Mises, Gamma, Poisson, Weibull, negative binomial, exponential, beta, log-normal, multinomial etc.). There certainly are no conceptual problems in using non-Gaussian mixtures for $\ln k_J(x, a_J)$, at the cost of programming analytical gradients (finite difference optimization is computationally very inefficient) and adjusting accordingly the computation of initial values. Depending upon the situation, we can also use alternative quadrature rules, such as Generalized Laguerre on $(0, \infty)$. Note, in particular, that the sparse grid approach discussed above allows for combining different types of univariate quadrature rules.

All in all, the algorithm we present in this paper can be extended in a number of ways to improve its flexibility at the cost of conceptually fairly straightforward though somewhat tedious additional programming.

5 Summary

We have proposed a generic sequential algorithm to construct Gaussian mixture approximations to analytically intractable density kernels. Our algorithm aims at minimizing a distance measure between the target kernel and the mixture that approximates the Monte Carlo variance of the corresponding IS ratio. In order to identify low probability terms, it currently relies upon products of univariate quadrature rules as an alternative to importance sampling. It is operational for low dimensions (say, up to three) but we expect to be able to handle higher dimensional targets by using instead sparse grid rules. For minimization of the distance measure we rely upon a quasi-Newton method using analytical gradient. Reliance upon analytical gradients requires one-time programming under an appropriate parametrization but has proved computationally much more efficient than minimizers relying upon finite difference or simplex optimizers. Extensions to other mixture types are computationally straightforward at the cost of programming of the corresponding gradients and adjusting accordingly the computation of initial values for the mixture terms. Pilot applications have demonstrated the flexibility as well as numerical accuracy of our algorithm.

Foremost, it is applicable to a wide range of important empirical mixture applications of considerable interest in the statistical and econometric literature. Two such applications are currently under development. One consists of a mixture filtering extension of the Kalman filter applicable to a broad range of dynamic state-space models combining a linear Gaussian latent fields with non-linear or non-Gaussian measurement densities. Essentially, the Kalman filter swarms of particles (mixtures of Dirac measures) are replaced by sequential finite Gaussian mixtures. The other application aims at producing finite mixture approximations to nonparametric density kernels. By reducing the number of terms well below the number of data points, we aim at facilitating the interpretations of the result e.g. by identifying data clusters captured by individual mixture terms. Pilot applications have already proved highly promising.

Programs for our current algorithm are available at `http://sf.cbs.dk/nk`. Further developments will be added as they became available.

Appendix

The distance measure $f_J(a_J)$ in Equation (8) can be approximated by Equation (14), which we reproduce here:

$$\hat{f}_J(a_J) = \frac{1}{2}\sum_{i=1}^{N} w_i \left[\ln \varphi(x_i) - \ln k_J(x_i, a_J)\right]^2 . \tag{46}$$

In order to minimize $\hat{f}_J(a_J)$, we first need to adopt a parametrization that guarantees the positivity of the diagonal elements r^j_{ss} of the lower triangular Cholesky factor R_j. This is achieved by re-parameterizing r^j_{ss} as $\exp\{\tilde{r}^j_{ss}\}$. Hence, the set of auxiliary parameters consists of $(\mu_j, \{r^j_{ts}\}_{t<s}, \{\tilde{r}^j_{ss}\}, \delta_j)$. The gradient of $\hat{f}_J(a_J)$ with respect to $(\mu_j, \{r^j_{ts}\}_{t<s}, \{\tilde{r}^j_{ss}\}, \delta_j)$ is given by

$$g = \sum_{i=1}^{N} w_i \frac{[\ln(\varphi(x_i)) - \ln k_J(x_i, a_J)]}{k_J(x_i, a_J)} \sum_{h=1}^{J} e^{\delta_h} k(x_i, \alpha_h) d_h(x_i), \tag{47}$$

where the summation in h represents the gradient of $k_J(x_i, a_J)$ with respect to $(\mu_j, \{r^j_{ts}\}_{t<s}, \{\tilde{r}^j_{ss}\}, \delta_j)$. The vector $d_h(x_i)$ consists of the following components

$$\begin{aligned}
&d^{\mu}_h(x) = R_h R_h^{'}(x - \mu^h)\\
&d^{r}_{tsh}(x) = -(x_s - \mu^h_s) e_t R_h^{'}(x - \mu^h) \text{ if } t < s \text{ for } t, s = 1, ..., d\\
&d^{\tilde{r}}_{ssh}(x) = -(x_s - \mu^h_s) e_s R_h^{'}(x - \mu^h) \exp\{\tilde{r}^h_{ss}\} + 1 \text{ for } s = 1, ..., d,\\
&d^{\delta}_h(x) = 1,
\end{aligned}$$

where e_s for $s = 1, ..., d$ is the d-dimensional vector, which consists of zeros and a unity at the s'th element of that vector, and μ^h_s is the s element of d-dimensional vector of means μ^h for $h = 1, ..., J$.

Acknowledgements The authors have benefited from discussions with Dave DeJong and Roman Liesenfeld.

References

Adcock CJ (2004) Capital asset pricing in uk stocks under the multivariate skew-normal distribution. In: Genton MG (ed) Skew-elliptical distributions and their applications, Chapman & Hall/CRC, London, pp 191–204

Adcock CJ (2010) Asset pricing and portfolio selection based on the multivariate extended skew-student-t distribution. Annals of Operations Research 176(1):221–234

Ardia D, Hoogerheide L, van Dijk H (2009) Adaptive mixture of student-t distributions as a flexible candidate distribution for efficient simulation: The r package admit. Journal of Statistical Software 29(1):1–32

Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726

Basturk N, Hoogerheide L, Opschoor A, van Dijk H (2012) The R package MitISEM: Mixture of student-t distributions using importance sampling weighted expectation maximization for efficient and robust simulation. Tinbergen Institute Discussion Paper 12-096/III, Tinbergen Institute

Bornkamp B (2011) Approximating probability densities by iterated laplace approximations. Journal of Computational and Graphical Statistics 20(3):656–669

Bradley PS, Fayyad UM (1998) Refining initial points for k-means clustering. In: Shavlik JW (ed) Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, USA, pp 91–99

Bungartz HJ, Griebel M (2004) Sparse grids. Acta Numerica 13:147–269

Cameron SV, Heckman JJ (2001) The dynamics of educational attainment for black, hispanic, and white males. Journal of Political Economy 109(3):455–499

Cappé O, Guillin A, Marin JM, Robert CP (2004) Population monte carlo. Journal of Computational and Graphical Statistics 13:907–929

Cappé O, Godsill SJ, Moulines E (2007) An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo. Proceedings of the IEEE 95(5):899–924

Cappé O, Douc R, Guillin A, Marin JM, Robert CP (2008) Adaptive importance sampling in general mixture classes. Statistics and Computing 18(4):447–459

DeSarbo WS, Degeratu AM, Wedel M, Saxton M (2001) The spatial representation of market information. Marketing Science 20(4):426–441

Domínguez-Molina J, González-Farías G, Ramos-Quiroga R (2004) Skew-normality in stochastic frontier analysis. In: Genton MG (ed) Skew-elliptical distributions and their applications, Chapman & Hall/CRC, London, pp 223–242

Domínguez-Molina J, González-Farías G, Ramos-Quiroga R, Gupta AK (2007) A matrix variate closed skew-normal distribution with applications to stochastic frontier analysis. Communications in Statistics - Theory and Methods 36(9):1691–1703

Douc R, Guillin A, Marin JM, Robert CP (2007) Minimum variance importance sampling via population monte carlo. ESAIM: Probability and Statistics 11:427–447

Doucet A, de Freitas N, Gordon N (2001) Sequential Monte Carlo Methods in Practice. Springer New York

Duffie D, Pan J (1997) An overview of value at risk. The Journal of Derivatives 4(3):7–49

Duong T (2007) ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. Journal of Statistical Software 21(1):1–16

Everitt BS, Hand DJ (1981) Finite Mixture Distributions. Chapman and Hall, London

Ferguson TS (1973) A bayesian analysis of some nonparametric problems. The Annals of Statistics 1(2):209–230

Ferreira JT, Steel MF (2007) Model comparison of coordinate-free multivariate skewed distributions with an application to stochastic frontiers. Journal of Econometrics 137(2):641 – 673

Frühwirth-Schnatter S (2006) Finite mixture and Markov Switching Models. Springer, New York

Geweke J (1996) Monte carlo simulation and numerical integration. In: Amman HM, Kendrick DA, Rust JP (eds) Handbook of Computational Economics, North Holland, Amsterdam, The Netherlands, pp 731–800

Gilks WR, Roberts GO, Sahu SK (1998) Adaptive markov chain monte carlo through regeneration. Journal of the American Statistical Association 93(443):1045–1054
Giordani P, Kohn R (2010) Adaptive independent metropolis-hastings by fast estimation of mixtures of normals. Journal of Computational and Graphical Statistics 19(2):243–259
Hamerly G, Elkan C (2002) Alternatives to the k-means algorithm that find better clusterings. In: Nicholas C, Grossman D, Kalpakis K, Qureshi S, van Dissel H, Seligman L (eds) Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM, New York, NY, USA, pp 600–607
Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57(2):357–384
Han B, Comaniciu D, Zhu Y, Davis LS (2008) Sequential kernel density approximation and its application to real-time visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(7):1186–1197
Heiss F, Winschel V (2008) Likelihood approximation by numerical integration on sparse grids. Journal of Econometrics 144(1):62 – 80
Holmes GK (1892) Measures of distribution. Publications of the American Statistical Association 3(18/19):141–157
Hoogerheide L, Opschoor A, van Dijk HK (2012) A class of adaptive importance sampling weighted {EM} algorithms for efficient and robust posterior and predictive simulation. Journal of Econometrics 171(2):101 – 120
Hoogerheide LF, Kaashoek JF, van Dijk HK (2007) On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. Journal of Econometrics 139(1):154 – 180
Hull J, White A (1998) Incorporating volatility updating into the historical simulation method for value-at-risk. Journal of Risk 1:5–19
Kasahara H, Shimotsu K (2015) Testing the number of components in normal mixture regression models. Journal of the American Statistical Association 110(512):1632–1645
Keane MP, Wolpin KI (1997) The Career Decisions of Young Men. Journal of Political Economy 105(3):473–522
Kim S, Shephard N, Chib S (1998) Stochastic volatility: Likelihood inference and comparison with arch models. Review of Economic Studies 65(3):361–393
Kon SJ (1984) Models of stock returns-a comparison. Journal of Finance 39(1):147–165
Kurtz N, Song J (2013) Cross-entropy-based adaptive importance sampling using gaussian mixture. Structural Safety 42:35 – 44
Liesenfeld R, Richard JF (2006) Classical and Bayesian Analysis of Univariate and Multivariate Stochastic Volatility Models. Econometric Reviews 25(2-3):335–360
Marchenko YV, Genton MG (2012) A heckman selection-t model. Journal of the American Statistical Association 107(497):304–317
Mazzuco S, Scarpa B (2015) Fitting age-specific fertility rates by a flexible generalized skew normal probability density function. Journal of the Royal Statistical Society Series A 178(1):187–203
McLachlan GJ, Peel D (2000) Finite mixture models. J. Wiley & Sons, New York
Moe WW, Fader PS (2002) Using advance purchase orders to forecast new product sales. Marketing Science 21(3):347–364
Newcomb S (1886) A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics 8(4):343–366
Ogundimu EO, Hutton JL (2016) A sample selection model with skew-normal distribution. Scandinavian Journal of Statistics 43(1):172–190
Oh MS, Berger JO (1993) Integration of multimodal functions by monte carlo importance sampling. Journal of the American Statistical Association 88(422):450–456
Omori Y, Chib S, Shephard N, Nakajima J (2007) Stochastic volatility with leverage: Fast and efficient likelihood inference. Journal of Econometrics 140(2):425 – 449
Pearson K (1894) Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 185:71–110
Pitt MK, Shephard N (1999) Filtering via simulation: Auxiliary particle filters. Journal of the American Statistical Association 94(446):590–599

Richard JF, Zhang W (2007) Efficient High-Dimensional Importance Sampling. Journal of Econometrics 141(2):1385–1411

Ristic B, Arulampalam S, Gordon N (2004) Beyond the Kalman filter: particle filters for tracking applications. Artech House, Boston, London

Scott DW (1992) Multivariate Density Estimation. Wiley, New Jersey

Scott DW, Szewczyk WF (2001) From kernels to mixtures. Technometrics 43(3):323–335

Smolyak SA (1963) Quadrature and interpolation formulas for tensor products of certain class of functions. Soviet Mathematics, Doklady 148(5):1042–1045

Stock JH, Watson MW (2007) Why has u.s. inflation become harder to forecast? Journal of Money, Credit and Banking 39(1):3–33

Titterington M, Smith AF, Makov UE (1985) Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester

Tucker AL (1992) A reexamination of finite- and infinite-variance distributions as models of daily stock returns. Journal of Business and Economic Statistics 10(1):73–81

Venkataraman S (1997) Value at risk for a mixture of normal distributions: the use of quasi-bayesian estimation techniques. Economic Perspectives (Mar):2–13

Wang X, Wang Y (2015) Nonparametric multivariate density estimation using mixtures. Statistics and Computing 25(2):349–364

Weldon WF (1892) Certain correlated variations in *Crangon vulgaris*. Proceedings of the Royal Society of London 51:1–21

Weldon WF (1893) On certain correlated variations in *Carcinus maenas*. Proceedings of the Royal Society of London 54:318–329

West M (1992) Modelling with mixtures. In: Bernardo JM, Berger JO, DeGroot MH, Smith AF (eds) Bayesian Statistics 4, Oxford University Press, Oxford, United Kingdom, pp 503–524