Electronic Journal of Statistics
Vol. 8 (2014) 355–389
ISSN: 1935-7524
DOI: 10.1214/14-EJS887

Variational Bayesian inference with Gaussian-mixture approximations

O. Zobay

*Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, United Kingdom**

Abstract: Variational Bayesian inference with a Gaussian posterior approximation provides an alternative to the more commonly employed factorization approach and enlarges the range of tractable distributions. In this paper, we propose an extension to the Gaussian approach which uses Gaussian mixtures as approximations. A general problem for variational inference with mixtures is posed by the calculation of the entropy term in the Kullback-Leibler distance, which becomes analytically intractable. We deal with this problem by using a simple lower bound for the entropy and imposing restrictions on the form of the Gaussian covariance matrix. In this way, efficient numerical calculations become possible. To illustrate the method, we discuss its application to an isotropic generalized normal target density, a non-Gaussian state space model, and the Bayesian lasso. For heavy-tailed distributions, the examples show that the mixture approach indeed leads to improved approximations in the sense of a reduced Kullback-Leibler distance. From a more practical point of view, mixtures can improve estimates of posterior marginal variances. Furthermore, they provide an initial estimate of posterior skewness which is not possible with single Gaussians. We also discuss general sufficient conditions under which mixtures are guaranteed to provide improvements over single-component approximations.

AMS 2000 subject classifications: Primary 62F15; secondary 62E17.
Keywords and phrases: Approximation methods, variational inference, normal mixtures, Bayesian lasso, state-space models.

Received March 2011.

1. Introduction

Recently, the Variational Bayes (VB) method has attracted growing interest as an alternative to Monte-Carlo integration in computational Bayesian inference (for reviews from the perspectives of different fields see, e.g., Opper and Saad (2001), Smidl and Quinn (2005), Bishop (2006), Wainwright and Jordan (2008), Ormerod and Wand (2010)). Together with the ease of application, the increasing popularity of this method is mainly due to the significant increase in computation speed which can be achieved for wide classes of statistical models. However, a serious impediment to an even more widespread use arises from the fact that VB is only approximate and that, in general, it is very difficult to assess

*Present address: MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, UK.

the quality of VB calculations without comparing to exact results. A major task for research on variational Bayesian techniques is therefore the development of methodology that allows us to improve VB results in a simple and flexible way. The purpose of the present paper is to contribute to this objective in the context of certain types of Variational Bayes calculations.

The basic idea of the VB method consists in approximating a target probability density p, for example a complicated Bayesian posterior, by a simpler one for which one can compute inferences more easily. To this end, one defines a set of functions $\mathcal{Q}$ within which one finds the approximation q by minimizing the Kullback-Leibler (KL) distance or relative entropy between the functions in $\mathcal{Q}$ and the target p, i.e.,

$$q = \operatorname*{argmin}_{q' \in \mathcal{Q}} \int q' \log \frac{q'}{p}. \tag{1}$$

For simplicity, the members of the set $\mathcal{Q}$ will be called trial functions in the following.

The quality of the approximation crucially depends on the choice of $\mathcal{Q}$. Most applications of the VB method make use of what we will call the factorized or "nonparametric" approach. In this case, $\mathcal{Q}$ is chosen as the set of all completely or partially factorized trial functions, e.g., $q(x_{1:n}) = \prod_{i=1}^{n} q_i(x_i)$ with $x_{1:n}$ shorthand for the parameter vector $(x_1, \ldots, x_n)$, with no restriction on the form of the factors besides normalization (the fully factorized case is also known as mean-field VB). In practice, this approach is then mainly applied to "conjugate-exponential" Bayesian target distributions p, for which the data likelihood is in the exponential family and priors are chosen as conjugate. In this case, the optimized VB factors turn out to belong to the same classes of distributions as the priors, and the minimization problem (1) can be solved efficiently by means of an iteration scheme that updates a finite vector of parameters. This method is attractive as it is simple and often computationally very fast. However, major drawbacks consist in the complete neglect of correlations between variables in different factors, and the restriction to certain types of probability models. Furthermore, it requires substantial effort to find tractable sets of trial functions that improve upon the factorized form.

Alternative to this nonparametric approach, one can also choose the trial functions to belong to a prespecified class of distributions[1], typically the multivariate normal family (Opper and Archambeau 2009). The optimization (1) is then with respect to the mean μ and covariance matrix Σ of the Gaussian and has to be carried out using multidimensional numerical minimization procedures such as the conjugate-gradient method. So far this approach has received much less attention than the nonparametric method. This might be attributable, at least in part, to the fact that in general, the size of the covariance matrix, and hence the number of variables to be optimized, grows quadratically with the length n of the parameter vector $x_{1:n}$ (Opper and Archambeau 2009). However, as also pointed out in Opper and Archambeau (2009), the target distribution

[1] One can of course also consider the mixed or "semiparametric" case, where some factors within a factorization approach are treated parametrically.

p may impose strong constraints on the optimal form of Σ which considerably reduce the number of free parameters. In this way, the problem can be alleviated significantly (alternatively, one can also restrict the form of Σ "by hand", if a reasonable guess is available).

An obvious advantage of the parametric method over the factorization approach lies in the fact that the former is able to describe correlations between variables. Furthermore, it is also applicable to statistical models outside the conjugate-exponential class. However, these advantages come at the cost of restricting the functional form of the marginals.

In order to improve the quality of VB results, the use of mixture distributions as trial functions might suggest itself as a natural strategy. A wide variety of probability distributions can efficiently be approximated by mixture distributions, and it is typically straightforward to compute inferences for them. Unfortunately, for mixtures the analytical calculation of the (differential) entropy $S = -\int q \ln q$, and thus of the KL distance, is intractable, in general. The solution of the optimization problem (1) therefore becomes more demanding. Jaakkola and Jordan (1998) have derived a variational lower bound for the entropy based on a factorization approach, but on the whole, the study of mixture trial distributions has not received much attention in the literature so far.

In the present paper, we consider the use of normal mixture distributions within a parametric VB approach. Our work is thus complementary to the proposal of Jaakkola and Jordan (1998). The computation of the entropy is dealt with by making use of an alternative lower-bound approximation and restricting the choice of covariance matrices for the mixture components. In this way, efficient numerical calculations become feasible. As this treatment is only concerned with the entropy term S in the KL distance, the procedure can be applied whenever the VB problem can be solved for a single-component Gaussian, and it requires an only modestly increased effort. Furthermore, while the method cannot be guaranteed to always lead to an improved solution, we expect it to do so in a large number of cases. In the paper, we discuss the application of this approach to a non-Gaussian state-space model and to the Bayesian lasso as examples of a "realistic" use of the method, and show that it can provide appreciable improvements under appropriate conditions. Note that "improvement" is to be understood in this context as a reduction in KL distance, in agreement with the overall VB strategy.

The paper is organized as follows. In Sec. 2 we first discuss some general qualitative considerations regarding the use of mixture distributions in variational Bayesian calculations. We then give a representation of the entropy S in terms of overlap contributions from mixture components, somewhat reminiscent of the inclusion-exclusion formulas of set and probability theory. The mathematical form of this representation motivates a simple lower bound to the mixture entropy S which allows for an approximate calculation of the KL divergence. Section 3 describes the implementation of this framework using mixtures of normal distributions with suitable restrictions on the choice of covariance matrices. In order to illustrate some aspects of the general qualitative behavior of the approach, we discuss a simple model problem with an isotropic target probability

density. Section 4 derives two simple criteria that allow us to check whether a mixture is guaranteed to improve upon a single-component approximation. The criteria are obtained by performing stability analyses of the single-component minimum in the KL divergence with respect to mixture trial functions. In Sec. 5, mixture VB is applied to the non-Gaussian state-space model mentioned above. The posterior distribution of the hidden variables is approximated with single- and multi-component Gaussian trial functions. Comparisons to Markov chain Monte-Carlo results show appreciable improvements in the calculation of the variances, thus justifying the use of mixture trial distributions. We also show that mixtures allow us to approximately calculate posterior skewness, which is not possible with single Gaussians. Section 6 provides an example with realistic data by applying the method to the Bayesian lasso. Finally, Sec. 7 gives a brief summary and some concluding remarks.

2. Variational approximations with mixtures

In this section, we first discuss some general aspects of VB approximations with mixture distributions. We then consider the representation of the entropy in terms of overlap contributions of mixture components. This motivates a simple lower-bound approximation of the mixture entropy which forms the basis of the subsequent developments. First of all, however, we return to the general formulation (1) of the VB optimization problem and recall two basic consequences that we will refer to later on.

2.1. General considerations

(i) We can think of the optimal VB solution as being determined by a competition between the negative-entropy part $\bar{S} := -S = \int q \ln q$ and the "energy term" $E = -\int q \ln p$ in the KL divergence. The minimization of the energy E alone would localize q as a Dirac delta function at the global maximum of the target distribution p. However, this contraction is counteracted by the influence of the entropy which tries to spread out the distribution q as much as possible. The actual minimum of the KL distance strikes a balance between these two opposing tendencies. Figure 1 shows a schematic graphical illustration of this principle. The horizontal coordinate of the graph represents a generic measure of the dispersion of the trial functions; for example it could stand for the width of a trial distribution centered at the maximum of p, or it might indicate the distance between two mixture components which are otherwise kept fixed.

(ii) The VB approximation q tends to be localized within the support of the target distribution as far as possible, i.e., typically q is small wherever p is, but not necessarily vice versa (Minka 2005). This is simply because otherwise the energy term $E = -\int q \ln p$ would impart a huge penalty on the KL divergence. As the trial functions cannot reproduce the shape of the target perfectly, this behavior results in the underestimation of variances as a typical feature of VB

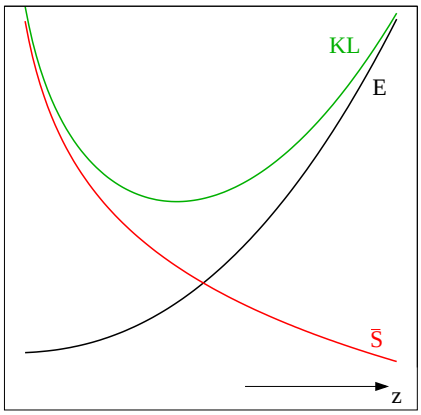

FIG 1. *Schematic showing typical dependence of negative entropy $\bar{S}$, energy E, and KL distance $\bar{S} + E$ on some generic coordinate z representing dispersion of trial functions (see text).*

approximations. For factorized VB, this problem increases with the strength of correlations between variables in the target (Bishop 2006).

We now turn to the discussion of VB approximations with mixture distributions, i.e., we assume the set $\mathcal{Q}$ to consist of trial functions of the form $\tilde{q} = \sum_{i=1}^{k} w_i \tilde{q}_i$ with $w_i \geq 0$, $\sum_{i=1}^{k} w_i = 1$. The mixture components $\tilde{q}_i$ are taken from some prespecified set $\mathcal{Q}_0$ of distributions. As schematically illustrated in Fig. 2, we expect there to be two main ways of how mixture distributions can improve the VB approximation. First of all, consider target distributions with multiple well-separated modes [Fig. 2(a)]. A single-component VB solution will typically be localized within one of the pieces of p, whereas an approximation with a mixture distribution may be able to describe the whole target. However, such situations can easily be handled with the help of a straightforward generalization of the one-component approach. Typically, the various parts of p will each give rise to a local minimum of the optimization problem, and the trial functions q_i corresponding to these minima will also be well-separated, i.e., have very small overlap. Under this crucial condition of negligible overlap, the q_i's are readily combined to a global approximation

$$q \propto \sum_i \exp(-K_i) q_i \tag{2}$$

of the target. In (2), K_i denotes the KL distance between q_i and p. An explanation and further discussion of this relation are given in Zobay (2009). Here we only emphasize that (2) shows how a careful analysis of the local solutions of the VB optimization problem (1) may already lead to efficient improvements of VB approximations in terms of mixtures. In such cases, one does not have to deal with complications arising from calculating the mixture's entropy.

Figure 2(b) shows the second and more interesting way of how mixtures can enhance VB approximations, which we will focus on in the rest of the paper. In

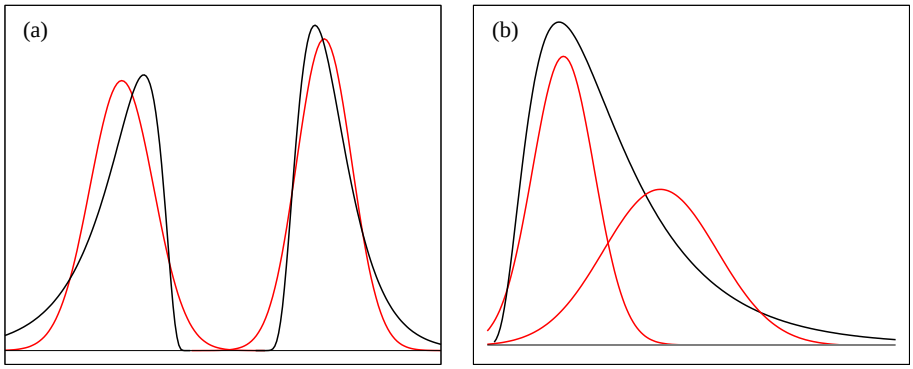

FIG 2. *Uses of mixture distributions in VB calculations. (a) Approximation of a multimodal target (black). Each mixture component (red) approximates a distinct piece of the target. (b) Mixture distribution providing improved description of shape of (unimodal) target.*

this situation, the mixture directly provides an improvement to the description of the form and shape of the (potentially unimodal) target distribution. In such cases, the single-component VB optimization problem may only have a unique solution, so that simple procedures in the spirit of (2) are not possible. The mixture approximation is thus genuinely different from the one-component case. The appropriate calculation of mixture entropies now becomes crucial, so that practical computations will be more complicated. It is difficult to predict a priori, without a detailed knowledge of the shape of the target, the extent of improvement that use of mixtures will provide. It is therefore important to develop methods that allow us to carry out such calculations with limited added effort, so that numerical results can be obtained quickly and easily.

Of course, one can imagine many situations which combine aspects of the two scenarios discussed above. One can consider, for example, a multimodal target as in Fig. 2(a) where each part is approximated by a mixture. The global solution is then obtained from (2) with the q_i representing the individual mixtures.

One may argue that a general drawback of the use of mixture distributions in VB is given by the fact that the potential improvement, as measured by the reduction of the KL distance, is strictly limited a priori. A k-component mixture can decrease the KL distance to the target distribution by an amount of at most $\log k$ compared to the optimal single-component approximation in $\mathcal{Q}_0$ (Jaakkola and Jordan 1998). However, one should keep in mind that the reduction in the KL distance gives only a very restricted view of the actual enhancement in the description of the target distribution. For example, in the case of a target distribution with two well-separated modes as shown schematically in Fig. 2(a), a single-component VB approximation is usually localized around one of the two modes, whereas a two-component VB solution will be able to represent the entire target. In this way, the description is improved substantially although the gain in the KL distance is at most $\log 2$. The example of Sec. 3.2 will also show that the approximation of the variance of a unimodal distribution can

be enhanced significantly although the change in the KL distance is relatively small, and similar conclusions may be drawn from the results of Sec. 5.

2.2. Entropy of a mixture distribution

For a mixture distribution $q = \sum_{i=1}^{k} w_i q_i$ with non-overlapping components (i.e., $w_i q_i(x) > 0$ for given x implies $w_j q_j(x) = 0$, $j \neq i$), the entropy $S = -\int q \ln q$ can be decomposed into the individual contributions of all mixture components, i.e.,

$$S = -\sum_{i=1}^{k} \int w_i q_i \ln(w_i q_i) = \sum_{i=1}^{k} S_1[w_i q_i]. \tag{3}$$

with $S_1[q] = -\int q \ln q$.

For the general case of overlapping mixture components, an extension of this representation can be derived. In addition to the contributions $S_1[w_i q_i]$ of the individual components, it also contains corrections due to the overlap between component distributions. These corrections are expressed as an expansion in terms of all possible combinations of two, three, or more components. For $k = 2$, a straightforward calculation shows that S can be written as

$$\begin{aligned} S &= -\int (w_1 q_1 + w_2 q_2) \ln (w_1 q_1 + w_2 q_2) = -\int w_1 q_1 \ln (w_1 q_1) \\ &\quad -\int w_2 q_2 \ln (w_2 q_2) - \int w_1 q_1 \ln \left(1 + \frac{w_2 q_2}{w_1 q_1}\right) - \int w_2 q_2 \ln \left(1 + \frac{w_1 q_1}{w_2 q_2}\right) \\ &= S_1[w_1 q_1] + S_1[w_2 q_2] + S_2[w_1 q_1, w_2 q_2] \end{aligned} \tag{4}$$

with S_2 given by the last two terms in the second line of (4). We can in fact interpret S_2 as describing the correction to the total entropy arising from the overlap of the two components, since the relevant integrands in the second line of (4) vanish at any argument x for which either $w_1 q_1(x)$ or $w_2 q_2(x)$ vanishes[2]. We also see that S_2 is always less than or equal to zero. Qualitatively, this is because an overlap between the mixture components reduces the overall uncertainty and hence the total entropy. Regarding the VB approximation, this implies that the entropic part of the KL divergence will give the mixture components a tendency to repel each other, as this will increase the entropy and hence lower the KL divergence.

For a three-component mixture, direct computation shows that the entropy is given by

$$S = \sum_{i=1}^{3} S_1[r_i] + S_2[r_1, r_2] + S_2[r_1, r_3] + S_2[r_2, r_3] + S_3[r_1, r_2, r_3] \tag{5}$$

[2]In the sense of $\lim_{x \to 0} x \log(1 + 1/x) = 0$.

with $r_i = w_i q_i$ and

$$S_3[r_1, r_2, r_3] = -\int r_1 \ln\left(1 - \frac{r_2}{r_1 + r_3}\frac{r_3}{r_1 + r_2}\right) - \int r_2 \ln\left(1 - \frac{r_1}{r_2 + r_3}\frac{r_3}{r_1 + r_2}\right) - \int r_3 \ln\left(1 - \frac{r_1}{r_2 + r_3}\frac{r_2}{r_1 + r_3}\right) \tag{6}$$

Similar to S_2, the term S_3 can be interpreted as giving the entropy contribution due to the common overlap of all three mixture components: the integrands in (6) vanish at any x for which any of the $r_i(x)$ vanishes. In this sense, (5) provides a decomposition of the total entropy into the contributions of the individual components, the overlaps of all pairs of components, and the common overlap of all three components. Relation (6) shows that S_3 is always positive. This can be explained by the fact that the addition of all pair overlaps S_2 overcounts the contribution of the three-component overlap which is subsequently corrected by S_3. In this way, (5) is already reminiscent of the well-known inclusion-exclusion formulas of set and probability theory.

Analogous behavior is found for the entropy expansion of four-component mixtures, suggesting the conjecture that it also holds in the general case. For $k = 4$, the total entropy is obtained as sum of the individual contributions, all possible two- and three-component overlap contributions as described above and the four-component overlap $S_4[r_1, r_2, r_3, r_4]$ given by the sum of

$$-\int r_1 \ln\left[1 + \frac{r_2 r_3 r_4}{(r_1 + r_2 + r_3)(r_1 + r_2 + r_4)(r_1 + r_3 + r_4)}\frac{(2r_1 + r_2 + r_3 + r_4)}{r_1}\right]$$

and the terms resulting from the cyclic permutations of the r_i's in the above expression. Due to the appearance of products such as $r_2 r_3 r_4$, the interpretation as overlap contribution is justified. In agreement with the inclusion-exclusion principle, S_4 is negative. However, as complete entropy expansions for $k \geq 4$ will not be needed in the following, the general case will not be investigated further here.

The above discussion suggests that a simple approximation to the total entropy of the mixture can be obtained by retaining only the individual and pairwise contributions S_1 and S_2 in the expansion. In fact, this provides a lower bound to S.

Proposition. *For a mixture distribution* $q = \sum_{i=1}^{k} w_i q_i$,

$$S[q] \geq \sum_{i=1}^{k} S_1[w_i q_i] + \sum_{i<j} S_2[w_i q_i, w_j q_j] =: -W[q] \tag{7}$$

with $S[q] = S_1[q] = -\int q \ln q$ *and* $S_2[q_1, q_2]$ *as defined in (4).*

Proof. By induction. For $k = 3$, the statement follows directly from (5) and (6).

Equivalently, we can write

$$\begin{aligned}\ln(r_1+r_2+r_3) &= \ln r_1+\ln\left(1+\frac{r_2}{r_1}\right)+\ln\left(1+\frac{r_3}{r_1}\right)\\ &\quad+\ln\left(1-\frac{r_2}{r_1+r_3}\frac{r_3}{r_1+r_2}\right)\\ &\leq \ln r_1+\ln\left(1+\frac{r_2}{r_1}\right)+\ln\left(1+\frac{r_3}{r_1}\right). \qquad (8)\end{aligned}$$

Multiplying (8) by $-r_1$, deriving corresponding inequalities for r_2 and r_3, and adding them up also provides the statement for $k=3$. The induction step follows in a similar way from

$$\begin{aligned}\ln(r_1+r_2+\cdots+r_k) &\leq \ln r_1+\ln\left(1+\frac{r_2+\cdots+r_{k-1}}{r_1}\right)+\ln\left(1+\frac{r_k}{r_1}\right)\\ &= \ln(r_1+r_2+\cdots+r_{k-1})+\ln\left(1+\frac{r_k}{r_1}\right)\\ &\leq \ln r_1+\ln\left(1+\frac{r_2}{r_1}\right)+\cdots+\ln\left(1+\frac{r_k}{r_1}\right).\end{aligned}$$

Here, the first line follows with the help of (8), whereas in the third line the induction hypothesis was used. □

In the following, we will find the VB solution under the approximation (8) or, more formally, we will solve the minimization problem

$$q=\operatorname*{argmin}_{q'\in\mathcal{Q}}\left\{W[q']-\int q'\ln p\right\} \qquad (9)$$

with $W[q]$ defined in (7). Several reasons motivate the study of this modification.

(i) As $W[q]\geq -S[q]$, $W[q]-\int q\ln p$ provides an upper bound to the exact KL divergence. We can therefore expect the minimization problem (9) to have well-defined solutions.

(ii) Calculating the two-component overlap contributions may, under certain conditions, be easier than computing the full mixture entropy. In our case, we will consider multivariate normal distributions with certain restrictions on the choice of the covariance matrices. The overlap contributions can then be calculated efficiently with the help of two-dimensional Gaussian integrations.

(iii) The most important motivation is based on the following consideration. The approximation of $\bar{S}=-S$ by W is worst (i.e., $\Delta:=W-\bar{S}$ is largest) if all mixture components overlap strongly, i.e., have similar locations, shapes, and scales. However, it will become progressively better when the components move away from each other or start to differ in form or extension, since in these cases the higher-order overlap contributions will become small. Now, unless a single-component solution already provides a very good description of the target, one would indeed expect the components of a mixture solution of (1)

to vary in location and/or scale, as only in this way a closer approximation of the target can be achieved. It is therefore reasonable to expect that in many cases the solutions of the two minimization problems (1) and (9) will not be too dissimilar.

This expectation is indeed confirmed by the numerical example at the end of Sec. 3 (see Fig. 4). There, very good agreement agreement between the solutions for (1) and (9) is found for the case of three mixture components, even though it is not obvious a priori that higher-order overlap contributions should be negligible.

3. VB with normal mixtures

3.1. General approach

As explained in the Introduction, multivariate normal distributions are an obvious choice for parametric variational Bayesian calculations. It is straightforward to incorporate and analyze correlations between variables, the entropy has a closed form, and the energy term in the KL distance is often amenable to analytic or efficient numerical calculation. More specifically, for a trial function $\tilde{q} \sim \mathcal{N}(\mu, \Sigma)$ [with $\mathcal{N}(\mu, \Sigma)$ denoting a k-dimensional multivariate normal distribution with mean μ and covariance matrix Σ], the KL distance is given by

$$\int \tilde{q} \ln \tilde{q} - \int \tilde{q} \ln p = -\ln \sqrt{(2\pi e)^k \det \Sigma} + E(\mu, \Sigma).$$

The minimization of the KL distance as a function of μ and Σ can usually be accomplished with the help of standard numerical techniques, such as conjugate gradient methods. The (local) minima are characterized by the equations (Opper and Archambeau 2009)

$$\nabla_\mu E(\mu, \Sigma) = 0, \tag{10}$$

$$\Sigma^{-1} + 2\nabla_\Sigma E(\mu, \Sigma) = 0. \tag{11}$$

An important consequence of (11) consists in the fact that it may impose strong restrictions on the form of the optimized precision matrix Σ^{-1} (Opper and Archambeau 2009). Assume, for example, that $\ln p \sim \sum_i f_i(x_i, x_{i+1})$. The precision matrix Σ^{-1} will then be tridiagonal. Opper and Archambeau (2009) also give another example which concerns Gaussian processes. These restrictions can significantly reduce the dimension of the search space for the optimization problem. However, even if such strict limitations do not exist, one is still free to "manually" impose restrictions on the form of the covariance matrix if they are suggested by the nature of the problem.

We now turn to the discussion of VB approximations with Gaussian mixtures $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$. The evaluation of the energy term in the KL distance does not impose any additional difficulties compared to the single-component case; the entropy term, however, becomes analytically intractable and therefore

presents the major challenge for practical calculations. In fact, the computation of entropies for Gaussian mixtures is a problem of considerable current interest and practical significance (see, e.g., Goldberger, Gordon and Greenspan (2003), Hershey and Olsen (2007), Huber, Bailey, Durrant-Whyte and Hanebeck (2008), Chen, Hershey, Olsen and Yashchin (2008)). Various approaches have been proposed, for example Monte-Carlo integration and Taylor-expansion schemes. However, for the present purposes, these methods did not appear suitable after an initial evaluation. Monte-Carlo calculations of sufficient accuracy are expected to be too slow for practical applications. Proposed deterministic approximations do not lead to an upper bound for the KL divergence, or their accuracy is difficult to judge.

Rather, to make the problem tractable, we have adopted the following strategy. First of all, we consider the approximate VB problem (9) which has been motivated in Sec. 2.2 and which amounts to replacing the exact entropy by single-component and pairwise contributions. However, even with this replacement the problem remains analytically intractable. We therefore stipulate that all covariance matrices Σ_i are multiples of a "base matrix" Σ_0, i.e., $\Sigma_i = \lambda_i^2 \Sigma_0$. In other words, we consider mixtures of "shape- and orientation-locked" normal distributions.[3] The calculation of the pairwise entropy then only requires a two-dimensional numerical integration which can be carried out efficiently by Gaussian quadrature. Before outlining the corresponding computations, however, we first give some comments on this strategy.

(i) In the single-component case, we have to calculate the mean μ_1 and the covariance matrix Σ_1 of the Gaussian. For a k-component mixture, we need $k-1$ additional mean vectors μ_i, $k-1$ weights w_i, and $k-1$ parameters λ_i. A single computation of the approximate KL distance requires $k-1$ additional energy and single-component entropy computations, as well as $k(k-1)/2$ pairwise entropy calculations. Since VB calculations are often significantly faster than MCMC, the increase in computational burden should be acceptable in most cases, as long as k does not become too large. As shown in the numerical examples, considerable improvements to single-component VB can be already be obtained for k as small as 2 or 3.

(ii) For mixtures of Gaussians, the limitations on the form of the covariance matrix implied by (11) no longer apply. In the example of Sec. 5, however, it was found that the base matrix Σ_0 still obeyed the restrictions of the single-component case to a very good degree of approximation. For practical purposes it might therefore be a reasonable strategy to first start with a restricted Σ_0 and then gradually relax the restrictions to see how much the results change.

(iii) Regarding the general approximation (9) to the VB problem we note that for a two-component mixture ($k = 2$), the approach still coincides with the exact VB optimization. For $k > 2$, $W + E$ provides an upper bound to the exact KL distance. Thus, any trial mixture with $k > 2$ for which $W + E$ is less than the minimum found for $k = 2$ is guaranteed to provide an improvement

[3]Note that if the normal mixture components separate into clusters such that normals in different cluster have negligible overlap [for example with a target as in Fig. 2(a)] we need locking only within clusters.

compared to the two-component case, even in terms of the exact KL distance. Note, however, that when we compare optimized trial functions with more than two components, we can of course no longer decide which one provides the best approximation in the sense of minimum KL divergence. For the numerical example of Sec. 5 we therefore focus on $k = 3$, although mixtures with more components could have been studied as well. Mixtures with $k > 2$ contain two-component distributions as a special case ($w_i = 0$ for $i > 2$) for which $W + E$ still provides the exact KL distance. It is therefore not unreasonable to assume that the solution of (9) for $k > 2$ can indeed improve upon the two-component case, and this has been observed in the example of Sec. 5.

(iv) The locking condition on the covariance matrices is of course a severe restriction which has only been imposed to make the problem tractable. One can easily imagine situations in which mixtures constructed in this way will not enhance the approximation significantly. In other cases, however, appreciable improvements may be obtained. Nevertheless, as argued above, the additional computational burden is modest and the result will, in any case, be at least as good as a single-component approximation. We therefore expect that the suggested approach should in many cases provide a viable option to improve upon standard VB results in a simple way. It should also be noted that in actual computations for concrete models one essentially only needs the calculation of the energy term as a plugin, as the entropy part is independent of the specific problem. In this sense, there is no additional effort compared to single-component calculations.

(v) We also note that mixtures allow us to estimate higher-order standardized moments, such as the skewness, which are inaccessible within a single-component approximation.

Details of the numerical calculation of the pairwise negative entropy $\bar{S}_2 = -S_2[w_1 q_1, w_2 q_2]$ are given in Appendix A. There it is shown that $\bar{S}_2$ effectively depends only on λ, the Mahalanobis distance $r^2 = (\mu_1 - \mu_2)^T \Sigma_1^{-1} (\mu_1 - \mu_2)$, and (up to a scaling factor) the ratio of the weights w_2/w_1. The behavior of $\bar{S}_2$ is depicted in Fig. 3. The curves illustrate the repulsion effect between mixture components which is due to the overlap entropy and which was mentioned in the discussion of (4). This can be seen as follows. Suppose the one-component VB problem is solved by a normal distribtion q_0. In the absence of $\bar{S}_2$, the two-component problem would then be solved by the "mixture" $\bar{q} = \frac{1}{2} q_0 + \frac{1}{2} q_0$ with $\lambda = 1$ and $r = 0$. However, as can be seen from Fig. 3, $\bar{S}_2$ is maximum for this mixture. The presence of $\bar{S}_2$ in the full two-component problem will thus tend to destabilize $\bar{q}$ as minimum of the KL divergence and to make the mixture components different from each other (however, in some cases $\bar{q}$ may still be the minimum of the full problem, depending on the shape of the KL landscape). Very importantly, Fig. 3(a) also shows that for growing k, $\bar{S}_2$ becomes more localized around $\lambda = 1$ (similar behavior is also found in the r dependence). This indicates that, in general, the differences between single- and multi-component VB solutions will be most pronounced (and hence our approach most useful) when the dimensionality of the problem is not too large.

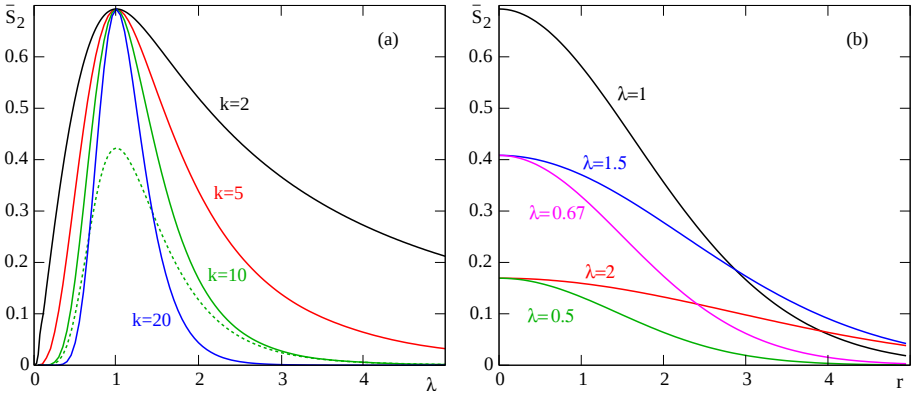

FIG 3. *Behavior of negative pairwise entropy* $\bar{S}_2 = -S_2$ *[equation (A.5)]. (a) Dependence on* λ *for* $r = 0$, $w_1 = w_2 = 0.5$, *and various* k. *Dashed line shows* $\bar{S}_2$ *for* $w_1 = 0.15$, $w_2 = 0.85$. *(b) Dependence on* r *for* $w_1 = w_2 = 0.5$, $k = 10$, *and various* λ. *Curve for* $\lambda = 1$ *is valid for all* k.

3.2. Example: Multivariate generalized normal distribution

To give a first illustration of the use of mixtures in VB, we now discuss a simple model problem. More elaborate and realistic applications will be studied in Secs. 5 and 6. In the present example, we consider the isotropic d-dimensional target density

$$p(x) = \frac{\beta\Gamma(d/2)}{2\pi^{d/2}\Gamma(d/\beta)} \exp\left(-|x|^\beta\right), \tag{12}$$

i.e., a multivariate generalized normal distribution with shape parameter (exponent) β. The exact expression for the variance of p is given by

$$\sigma_{\text{ex}}^2 = \frac{\Gamma\left(\frac{d+2}{\beta}\right)}{d\Gamma(d/\beta)}.$$

Due to the rotational symmetry of the target, it is a natural choice to use isotropic normal distributions $\mathcal{N}(0, \sigma_i^2 \mathrm{I})$ with variances σ_i^2, $1 \le i \le k$, in our k-component mixtures. In this way, the locking condition on the covariance matrices is fulfilled automatically. As an additional advantage, for any k we can compute the entropy of the mixture exactly in terms of one-dimensional quadratures. One can therefore easily study the effect of the approximation (9) to the original VB problem (1).

For the target (12), Fig. 4 shows the results of VB calculations with single normal distributions (black curves), two-component mixtures (red), and three-component mixtures with the entropy treated exactly (green) and approximately (blue). We focus on the VB variances (i.e., the variances of the solutions to the VB optimization problems) which might be considered the most interesting quantities in this context from the point of view of Bayesian inference. Figure 4

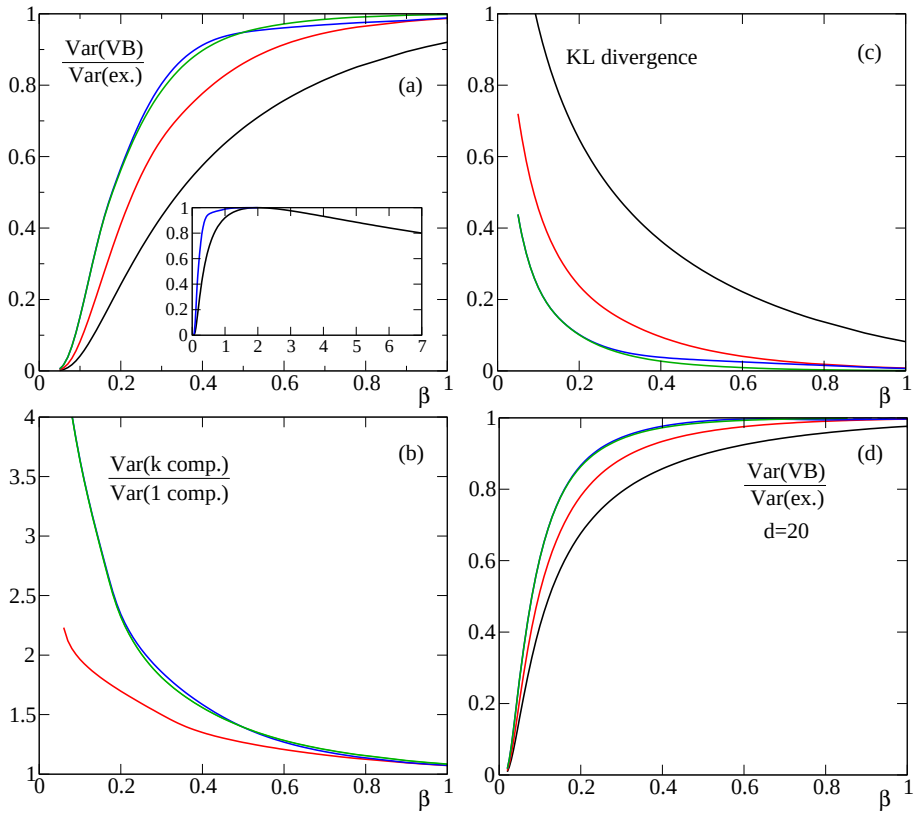

FIG 4. *Variational approximation of isotropic generalized normal density (12). (a) Ratio of variationally approximated variances to exact variance as a function of exponent β for $d = 5-$dimensional distribution. VB approximations use k-component normal mixtures with $k = 1$ (black curve), $k = 2$ (red), $k = 3$ (green), $k = 3$ and approximation (9) (blue). Inset shows results for a larger range of β. (b) Ratio of VB variances for two- and three-component mixtures to VB variance with single Gaussian. (c) Kullback-Leibler divergences. (d) Ratio of VB variances to exact variance for $d = 20$-dimensional distribution. Color coding of curves in (b)-(d) corresponds to (a).*

illustrates their dependence on the shape parameter β and the dimension d of the target. In addition, the KL distance is shown which can be computed exactly as the normalization constant of p is known explicitly.

For the single Gaussian, the VB problem can be solved analytically, yielding the variance

$$\sigma_1^2 = \left(\frac{d\Gamma(d/2)}{2^{\beta/2}\Gamma\left(\frac{d+\beta}{2}\right)\beta} \right)^{2/\beta} . \tag{13}$$

In the other cases, the optimization has to be performed numerically. The overall behavior of the VB approximation is illustrated in the inset of Fig. 4(a) which

shows the ratio $\sigma_1^2/\sigma_{\rm ex}^2$ as a function of β for a $d = 5$-dimensional target. For $\beta = 2$ (Gaussian target), VB is exact, but becomes less accurate the more β deviates from this value. For $\beta > 2$, i.e., for a target with light tails, it turns out that the VB approximations with two- and three-component mixtures coincide with the result from the single Gaussian.

However, for $\beta < 2$, i.e., heavy-tailed targets, a very different behavior is observed. Figures 4(a) and (b), respectively, show the ratio of the various VB variances to the true value $\sigma_{\rm ex}^2$, and the ratio of the variances for the two- and three-component mixtures to σ_1^2. We see that the use of mixture trial functions can significantly improve the variance estimates. This enhancement is particularly pronounced for smaller values of β, where the single-Gaussian approximation becomes less and less accurate (of course, for β close to 2, single Gaussian VB becomes better so that there is less scope for improvements by using mixtures). The KL divergences shown in Fig. 4(c) behave as expected from Fig. 4(a),(b) [blue curve shows $W + E$ of (9)]. The different types of behavior for $\beta < 2$ and $\beta > 2$, respectively, will be discussed further in Sec. 4 to exemplify the general stability criteria for single-component approximations derived there.

Figure 4(d) illustrates some effects of dimensionality. One sees that for increasing d, the single-Gaussian approximation to the variance is improving. This is because for larger d the probability mass of the target tends to be localized in a shell further away from the origin, as can be seen from the behavior of the radial probability density $p_r(r) \sim r^{d-1}p(r)$ with $r = |x|$. The distribution can therefore more easily be approximated by a Gaussian. The two- and three-component mixtures still provide clear improvements to the single Gaussian. However, the effect is somewhat reduced in the sense that for a given value of $\sigma_1^2/\sigma_{\rm ex}^2$, the additional improvement from the mixtures becomes smaller for growing d. We might perhaps attribute this effect to the increasing localization of the pairwise entropies $\bar{S}_2$ discussed in connection with Fig. 3.

It is also instructive to compare the three-component solutions for the original and the approximated VB problems (1) and (9). First of all, Fig. 4(c) shows that the approximate KL divergence $W + E$ (blue curve) is always less than the two-component KL distance (red). As discussed above, we can therefore be sure that the approximate three-component solution improves upon the two-component result in the KL sense. Furthermore, we also find the KL distance for the exact three-component calculation to be less than the approximate one, as it should. Nevertheless, all four diagrams show that the differences between the exact and approximate results are very small. The neglect of the higher-order overlap contribution to the entropy is thus well justified for this example. This result is remarkable because all mixture components are centered at 0, i.e., in the present case any effects of higher-order overlap contributions should be particularly pronounced. Judging from the mixture compositions, it is also not obvious that such contributions should be negligible. To give an example, at $\beta = 0.5$, the component have variance ratios $\sigma_1^2 : \sigma_2^2 : \sigma_3^2 = 0.2 : 1 : 4$ ($0.46 : 1 : 2.1$ for $d = 20$) and weights $0.24 : 0.53 : 0.23$ ($0.23 : 0.54 : 0.23$). Based these considerations and the numerical results, we can expect that (9) will also be a useful approximation in other cases.

4. Sufficient conditions for improvement of VB approximations by mixtures

In this section, we present two simple criteria that allow us to check whether a mixture VB approximation will be guaranteed to provide an improvement compared to the single-component case. These criteria are derived from the stability analysis of the single-component minima in the KL distance with respect to mixture trial functions.

We assume that the single-component minimizer of the KL divergence is given by the trial function $q(\theta_0)$ with parameter vector θ_0. There are two ways $q(\theta_0)$ can be "perturbed" by extending the space of trial functions to include mixture distributions: (a) by the addition of an arbitrary second component $q(\theta_1)$ with small weight, and (b) by slightly modifying the component parameters θ_i around θ_0 in a mixture with arbitrary weights. The first possibility corresponds to a stability analysis with trial functions $(1-\delta w)q(\theta_0)+\delta w q(\theta_1)$ and $\delta w \ll 1$, whereas for the second option trial functions of the form $\sum_i w_i q(\theta_0+\delta\theta_i)$ with small $\delta\theta_i$ are considered. We will discuss both situations in turn and derive corresponding criteria for the instability of the single-component solution.

In order to examine the first case, we consider the mixture $(1-w)q(\theta_0)+wq(\theta_1)$. Its KL divergence is given by

$$\begin{aligned}\mathrm{KL}(w;\theta_0,\theta_1|p) &= \int [(1-w)q(\theta_0)+wq(\theta_1)]\log[(1-w)q(\theta_0)+wq(\theta_1)] \\ &\quad - \int [(1-w)q(\theta_0)+wq(\theta_1)]\log p\end{aligned}$$

with p the target distribution. Assuming exchangeability of integration and differentiation, the first two derivatives of the KL divergence with respect to w read

$$\partial_w \mathrm{KL} = \int [-q(\theta_0)+q(\theta_1)]\log[(1-w)q(\theta_0)+wq(\theta_1)] \tag{14}$$

$$- \int [-q(\theta_0)+q(\theta_1)]\log p, \tag{15}$$

$$\partial_{ww}\mathrm{KL} = \int \frac{[-q(\theta_0)+q(\theta_1)]^2}{(1-w)q(\theta_0)+wq(\theta_1)}. \tag{16}$$

Taking the limit $w \downarrow 0$, the (one-sided) derivative of the KL divergence at $w=0$ is found as

$$\partial_w \mathrm{KL}\,|_{w=0} = \mathrm{KL}(\theta_1|p) - [\mathrm{KL}(\theta_1|\theta_0)+\mathrm{KL}(\theta_0|p)]\,. \tag{17}$$

The instability of the single-component minimum of the KL divergence with respect to the admixture of component $q(\theta_1)$ is therefore guaranteed if expression (17) is negative. This condition can be checked very easily in practical calcula-

tions[4] and thus be used in the search for good starting values for the iterative optimization of the mixture approximation. Note that if the KL divergence were to fulfill a triangle inequality, (17) would always be non-positive. However, as this is not the case, (17) can be both negative and positive.

It should also be noted that the second derivative (15) is non-negative for all values of w. This implies that that for $\mathrm{KL}(\theta_0|p) < \mathrm{KL}(\theta_1|p)$ and positive first derivative (17) at $w = 0$, all mixtures $(1-w)q(\theta_0) + wq(\theta_1)$ have larger KL divergence than $q(\theta_0)$. This holds independent of whether $q(\theta_0)$ is the actual single-component minimizer.

For the example of Sec. 3.2 involving an isotropic generalized normal target distribution approximated by isotropic Gaussians, criterion (17) can be evaluated analytically. Choosing $q(\theta_0)$ as the single-component minimizer determined by (13), it is found that for $\beta < 2$ and any d, (17) is negative for any $q(\theta_1)$, i.e., $q(\theta_0)$ is always unstable against admixture of a second component. For $\beta > 2$, however, (17) is always positive. In particular, this implies that any mixture $(1-w)q(\theta_0) + wq(\theta_1)$ has a larger KL divergence than $q(\theta_0)$. Nevertheless, this argument does not rule out that mixture distributions not involving $q(\theta_0)$ may have a lower KL divergence. This can only be checked numerically.

We now turn to the second way of perturbing the single-component solution which involves modification of the component parameters. Let $\mathcal{Q}^{(1)}$ be a family of trial functions $q(\theta)$ which smoothly depend on the p-dimensional parameter vector θ. The KL distance $\mathrm{KL}^{(1)}(\theta)$ to the target function p and the first two derivatives of $\mathrm{KL}^{(1)}$ with respect to θ are given by

$$\mathrm{KL}^{(1)}(\theta) = \int q(\theta)\log q(\theta) - \int q(\theta)\log p, \tag{18}$$

$$\partial_\theta \mathrm{KL}^{(1)}(\theta) = \int \partial_\theta q(\theta)\log q(\theta) - \int \partial_\theta q(\theta)\log p, \tag{19}$$

$$\partial_{\theta\theta} \mathrm{KL}^{(1)}(\theta) = A(\theta) + B(\theta) \tag{20}$$

with the matrices $A(\theta)$ and $B(\theta)$ defined by

$$A(\theta) = \int \partial_{\theta\theta} q(\theta)\log q(\theta) - \int \partial_{\theta\theta} q(\theta)\log p, \tag{21}$$

$$B(\theta) = \int \frac{1}{q(\theta)}\partial_\theta q(\theta)[\partial_\theta q(\theta)]^T. \tag{22}$$

Let θ_0 be the global minimizer of $\mathrm{KL}^{(1)}(\theta)$, assumed to lie within the interior of the parameter space. This implies that $\partial_\theta \mathrm{KL}^{(1)}(\theta_0) = 0$. We assume that the matrix of second derivatives $\partial_{\theta\theta}\mathrm{KL}^{(1)}(\theta_0)$ is positive definite, thus excluding some exceptional cases.

[4] The difference $\mathrm{KL}(\theta_1|p) - \mathrm{KL}(\theta_0|p)$ is available from the single-component computations. In case p is of the form $p(x|y) = p(x,y)p(y)$, the potentially problematic evidence terms $-\log p(y)$ in the KL divergence cancel. The KL divergence $\mathrm{KL}(\theta_1|\theta_0)$ can be derived analytically for Gaussian trial functions.

We now consider the family $\mathcal{Q}^{(k)}$ of k-component mixture distributions derived from $\mathcal{Q}^{(1)}$ and characterized by parameters $(\theta^{(1)}, \ldots, \theta^{(k)})$ and weights $w_1, \ldots, w_k$. The set $\mathcal{Q}_0^{(k)}$ of functions within $\mathcal{Q}^{(k)}$ that are equivalent to the single-component minimizer $q(\theta_0)$ have constrained parameters $(\theta^{(1)} = \theta_0, \ldots, \theta^{(k)} = \theta_0)$ but unrestricted weights $\sum_i w_i = 1$.

In the following, we first show that the functions in $\mathcal{Q}_0^{(k)}$ are stationary points of the k-component KL distance $\mathrm{KL}^{(k)}$. We then investigate the stability of these stationary points with respect to variations of the $\theta^{(i)}$. It turns out that the stability depends on the eigenvalues of the matrix $A(\theta)$. If A has negative eigenvalues, the stationary points will be unstable. This means that the global minimum of $\mathrm{KL}^{(k)}$ is less than $\mathrm{KL}^{(1)}(\theta_0)$, i.e., the mixture distributions are guaranteed to lead to an improved approximation compared to the single-component case. If all eigenvalues are strictly positive, the k-component stationary points will be stable, i.e., they form a local minimum. In this case, the local stability analysis cannot predict anything about the possible improvements by mixtures, as we cannot decide whether these local minima are also global. In the marginal case of A having positive and zero eigenvalues, we also cannot draw any conclusions. Note that this discussion is independent of the number of mixture components.

More formally, these results are summarized as follows (proofs are provided in Appendix B).

Theorem 1. *The functions in $\mathcal{Q}_0^{(k)}$ are stationary points of of the KL distance $\mathrm{KL}^{(k)}$.*

The independence of $\mathrm{KL}^{(k)}$ on the w_i's also directly implies that $\partial_{w_j w_l}\mathrm{KL}^{(k)} = 0$ and $\partial_{\theta^{(j)}, w_l}\mathrm{KL}^{(k)} = 0$. The stability of the stationary points is therefore determined by the matrix $\partial_{\theta^{(j)},\theta^{(l)}}\mathrm{KL}^{(k)}$.

Theorem 2. *For the functions in $\mathcal{Q}_0^{(k)}$,*

$$\det\left[\partial_{\theta^{(j)},\theta^{(l)}}\mathrm{KL}^{(k)}\right] = (\det A)^{k-1}\det(A+B)\prod_{i=1}^{k} w_i \tag{23}$$

with A and B defined in (21) and (22) with $\theta = \theta_0$. At $w_1 = \cdots = w_k = 1/k$, the eigenvalues λ of $\partial_{\theta^{(j)},\theta^{(l)}}\mathrm{KL}^{(k)}$ are the roots of the characteristic equation

$$[\det(A - k\lambda 1_p)]^{k-1}\det(A + B - k\lambda 1_p) = 0 \tag{24}$$

with 1_p the p-dimensional identity matrix.

Corollary 1. *If A has negative eigenvalues, the global VB optimizer in the class $\mathcal{Q}^{(k)}$ of k-component mixture distributions is guaranteed to improve upon the best single-component approximation $q(\theta_0)$.*

Corollary 2. *If all eigenvalues of A are positive, the KL divergence $\mathrm{KL}^{(k)}$ has a local minimum at the k-component mixture distributions corresponding to the best single-component solution $q(\theta_0)$.*

To illustrate these results, we again apply them to the example discussed in Sec. 3.2. In this case, the parameter vector θ is one-dimensional and only contains the variance σ^2 of the isotropic Gaussian trial function. The determinant $\det(A(\theta_0))$ at the minimizer θ_0 of the single-component problem be calculated analytically. It is found to be proportional to $2 - \beta$, with β the shape parameter of the target. Corollaries 1 and 2 thus imply that for $\beta < 2$, mixture trial functions are guaranteed to provide an improvement, whereas for $\beta > 2$ the single-component solution remains a local minimum. These conclusions are in agreement with the results obtained from criterion (17) and are confirmed by the numerical calculations described in Sec. 3.2. In particular, for $\beta > 2$ the results suggest that the single-component solution is in fact the global minimum, as no better approximation could be found.

We conclude this section with a brief comment on the application of the criteria to the shape- and orientation-locked Gaussian mixture trial functions described in Sec. 3.1. Here one should keep in mind that the parameter vector θ cannot contain the unconstrained covariance matrix due to the locking condition. To simplify the practical use of the criteria, one could therefore consider mixture components $\mathcal{N}(\mu, \beta\Sigma_0)$ where only μ and the scalar β are variable parameters. The covariance matrix Σ_0 would be kept fixed, e.g., at the result for the single-component minimizer.

5. A non-Gaussian state-space model

Following the study of an illustrative example with an isotropic target density in Sec. 3, we now investigate a non-Gaussian linear state-space model as a more realistic application of the mixture method. The choice of this model is partly motivated by the results of Sec. 3.2, where we found that the use of mixtures can provide appreciable improvements for heavy-tailed target distributions. However, it should be emphasized that the approach can be applied to any model that is amenable to VB treatment with a single Gaussian. As is common with variational calculations, the potential benefits are very difficult to gauge a priori, and empirical studies are needed, in general.

Our non-Gaussian state-space model for the random variables $X_{1:n}$ and $Y_{1:n}$ is defined by

$$X_{i+1} = \phi X_i + \eta_i, \qquad \eta_i \sim \mathcal{GN}(0, \beta, \kappa), \tag{25}$$

$$Y_i = X_i + \xi_i, \qquad \xi_i \sim \mathcal{GN}(0, \alpha, \rho), \tag{26}$$

where we have set $X_0 \equiv 0$. Here, $\mathcal{GN}(\beta, \mu, \kappa)$ denotes the (univariate) generalized normal distribution with mean μ, shape and scale parameters β and κ, respectively, and density

$$p_{gn}(x; \mu, \beta, \kappa) = \frac{\beta}{2\kappa\Gamma(1/\beta)} \exp\left[-\left(\frac{|x-\mu|}{\kappa}\right)^{\beta}\right]. \tag{27}$$

The complete probability distribution for the model is thus given by

$$p(x_{1:n}, y_{1:n}) = \prod_{i=1}^{n} p_{gn}(x_i; \phi x_{i-1}, \beta, \kappa) p_{gn}(y_i; x_i, \alpha, \rho). \tag{28}$$

In the following, we will assume that the model parameters $(\alpha, \beta, \kappa, \rho, \phi)$ are known, and that data $y_{1:n}$ have been observed. The objective is then to study the posterior distribution $p(x_{1:n}|y_{1:n})$. In the VB treatment of this model, we will compare the approximation using a single Gaussian to two- and three-component mixtures. The three-component calculation is based on the modified VB optimization problem (9). Details of the variational calculations are given in Appendix C. We note that the dimensionality of the minimization problem scales linearly with the number of observations. To assess the overall accuracy of the variational results, we have also performed MCMC calculations using a standard Gibbs sampling approach. Draws from the corresponding univariate conditional densities are obtained using adaptive rejection Metropolis sampling (Gilks, Best and Tan 1995).

In the subsequent discussion, we will focus on the variational results for mean, variance, and skewness. In particular, it should be emphasized that the skewness (or any other higher-order moment) cannot be obtained from single-component VB, but becomes accessible through the mixture approach. In view of the fact that nonparametric variational calculations would be very difficult for the given model, we can thus conclude that Gaussian-mixture VB provides a genuine extension of the capabilities of variational calculations.

Our numerical studies of the model (25), (26) confirm some general trends which are already observed in the simple example of Sec. 3.2. It is found that the mixture approach is most effective for heavy-tailed distributions (i.e., α and β less than 2) and moderate dimensionality (i.e., number n of observations not too large).

In the subsequent numerical examples, we will therefore focus on the case of equal shape parameters $\alpha = \beta$ for which the values 1.2 and 0.35 are chosen in two sets of calculations. The other model parameters are kept fixed at $\kappa = 1$, $\rho = 0.2$, and $\phi = 0.5$. The observations $y_{1:n}$ are generated from the model (25), (26) with the given parameters and n set to 5 or 15. For each parameter setting, 200 random samples are investigated in order to get an overview of the performance of the approach.

The main findings can be summarized as follows. In all cases, mixture VB reduces the KL divergence, i.e., it improves the approximation of the posterior. MCMC means are reproduced quite accurately already with a single Gaussian. VB tends to underestimate variances, but the mixture approach increases single-component results by up to 30% on average depending on parameters, with larger increases for smaller samples and growing non-Gaussianity. The variational skewness estimates usually reproduce the sign of the MCMC results although they tend to underestimate the actual value.

For a more detailed discussion, we first turn to the case $\alpha = \beta = 1.2$ for which VB calculations of posterior variance and skewness are shown in Fig. 5

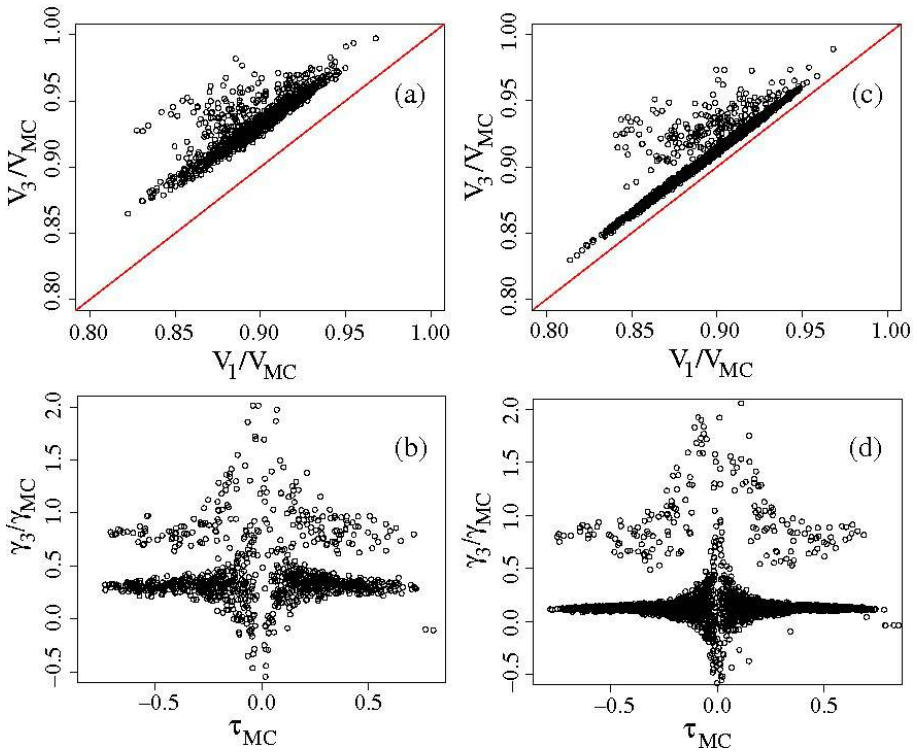

FIG 5. *Variational study of non-Gaussian state space model (25), (26) for parameters* $\alpha = \beta = 1.2$, $\kappa = 1$, $\rho = 0.2$, $\phi = 0.5$ *and number of observations* $n = 5$ *(a,b) and* $n = 15$ *(c,d). For both parameter settings,* 200 *random samples generated from the model were investigated. Figures (a) and (c) show the ratio* V_3/V_{MC} *of marginal variances calculated from three-component VB and MCMC as a function of* V_1/V_{MC} *with* V_1 *the variance from the single-component calculation. Red lines show diagonal* $y = x$*. These diagrams illustrate the change in the estimated variance when using mixture VB. Figures (b) and (c) depict the skewness ratio* γ_3/γ_{MC} *as a function of the standardized third moment* τ_{MC} *as obtained from MCMC.*

(means will be briefly mentioned in the discussion of Fig. 7 which refers to $\alpha = \beta = 0.35$). Results for the variance are shown in Figs. 5(a) ($n = 5$) and 5(c) ($n = 15$). The diagrams depict the variance ratio V_3/V_{MC} as a function of V_1/V_{MC}. Here, V_1 and V_3 denote the variance of a specific posterior marginal $p(x_i)$ as calculated from one- and three-component VB, whereas V_{MC} is the corresponding MCMC result. We see that the single-component results are already quite accurate as V_1/V_{MC} ranges from 0.85 to 0.95 (note that variances are underestimated as discussed in Sec. 2.1). All points in Figs. 5(a) and (c) lie between the diagonal $y = x$ (red line) and the horizontal $y = 1$. This means that in all cases $V_1 < V_3 < V_{MC}$, i.e., the three-component calculation always provides an improved estimate of the variance. As V_1 already yields a very good approximation, the improvement is not large, but nevertheless clearly obvious and significant. We also see that for growing n the average improvement is re-

duced, e.g., for $V_1/V_{MC} \approx 0.90$ we find V_3/V_{MC} to approximately range within 0.934 ± 0.005 for $n = 5$ and 0.912 ± 0.002 for $n = 15$.

The benefits of the mixture calculation become even more obvious when studying the results for the skewness as this quantity cannot be obtained from single-Gaussian VB. Skewness is defined as

$$\gamma = \frac{\tau}{V^{3/2}} \tag{29}$$

with V the variance and $\tau = \mathbb{E}[(X - \mathbb{E}(X))^3] = \mathbb{E}(X^3) - 3\mathbb{E}(X^2)\mathbb{E}(X) + 2\mathbb{E}(X)^3$ the third moment about the mean. In Figs. 5(b),(d) (for sample sizes $n = 5$ and $n = 15$, respectively), the VB-to-MCMC skewness ratio γ_3/γ_{MC} is shown as a function of the third moment τ_{MC} obtained from MCMC. We first of all find that the VB calculation provides a reliable estimate of the overall skew of the posterior (i.e., the sign of γ). In both calculations, one finds that in only about 5% of all cases the ratio γ_3/γ_{MC} is negative, i.e., γ_3 and γ_{MC} have opposing signs. As can be seen from the diagrams, these discrepancies predominantly occur when $|\tau_{MC}|$ is small. This behavior can be explained by the fact that for small $|\tau|$, the estimate of τ will be very sensitive to any errors in the (MCMC or variational) calculations of the moments $\mathbb{E}(X^j)$. For larger $|\tau_{MC}|$, most VB results appear to underestimate γ_{MC} by an almost constant factor (about 0.3 in (b) and 0.1 in (d)). However, there is also a distinct group of cases for which the ratio γ_3/γ_{MC} becomes close to one. Altogether, we thus find that mixture VB provides a useful first approximation of posterior skewness.

Results for the case of $\alpha = \beta = 0.35$ are displayed in Figs. 6 and 7. Figures 6(a),(b),(d),(e) show variance ratios for $n = 5$ and $n = 15$, with (b) and (e) providing enlargements of (a) and (d), respectively, for small V_1/V_{MC}. For $\alpha = \beta = 0.35$, the model exhibits much stronger non-Gaussianity than in the previous case and consequently the VB approximation becomes less accurate. As can be seen from Figs. 6(a) and (d), the variance ratio V_1/V_{MC} extends over a much larger range extending from 0 to 1 with most values, however, in the region of small V_1/V_{MC}. For small ratios V_1/V_{MC}, Figs. 6(b) and (e) indicate that mixture VB brings about a marked improvement in the estimation of the variances. We also see that for small V_1/V_{MC}, we have only a few cases with $V_3 < V_1$. However, this behavior changes as V_1/V_{MC} becomes larger as estimates with $V_3 < V_1$ occur more often.

More specifically, 66% of all cases for $n = 5$ (68% for $n = 15$) have $V_1/V_{MC} < 0.5$. For 92% (96%) of these cases, the mixture approximation improves the variance estimates, i.e., $V_3 > V_1$. The extent of improvement, as measured by mean of the variance ratio V_3/V_1, is 1.49 (1.13). When we consider all cases, the percentages decrease to 73% (83%) and the ratios to 1.31 (1.08). It thus appears as if the benefits of the mixture approximation are strongest where single-component VB performs badly (i.e., small V_1), at the expense of reduced improvement where single-component VB is good already. At any rate, it should be kept in mind that for every sample the KL divergence of the mixture approximation is lower than the single-component value, i.e., the overall approximation of the posterior is improved even if some variance estimates become somewhat less accurate.

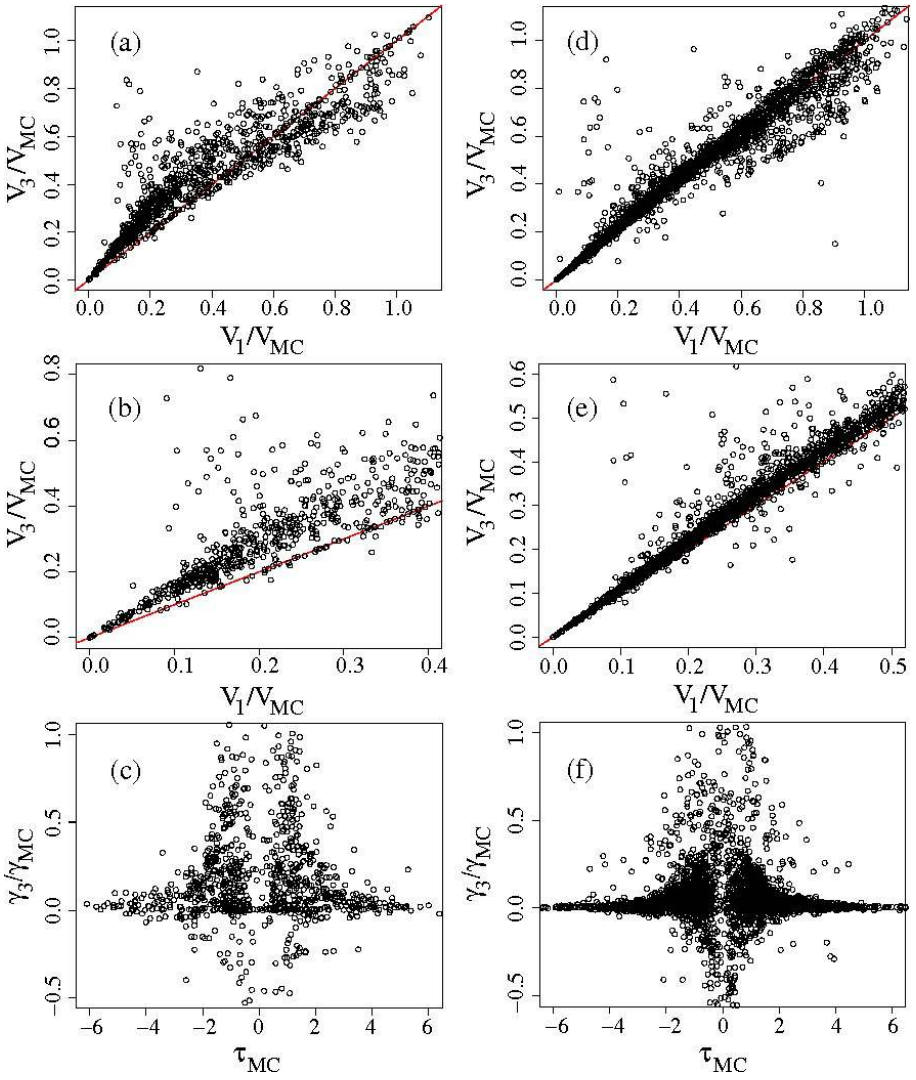

FIG 6. *Variational study of non-Gaussian state space model (25), (26). Parameters as in Fig. 5 except for* $\alpha = \beta = 0.35$. *Number of observations are* $n = 5$ *(a,b,c) and* $n = 15$ *(d,e,f). Diagrams show variance and skewness ratios obtained for 200 random samples analogous to Fig. 5. Figures (b) and (e) depict detail of (a) and (d), respectively, for small* V_1/V_{MC}.

The overall skew is still determined correctly in about 77% of all cases (i.e., γ_3 and γ_{MC} have the same sign). As can be expected, Figs. 6(c) and (f) show that the skewness ratios γ_3/γ_{MC} are lower than in the case of $\alpha = \beta = 1.2$. Altogether, the results show that the application of mixture VB still remains useful in the study of this model.

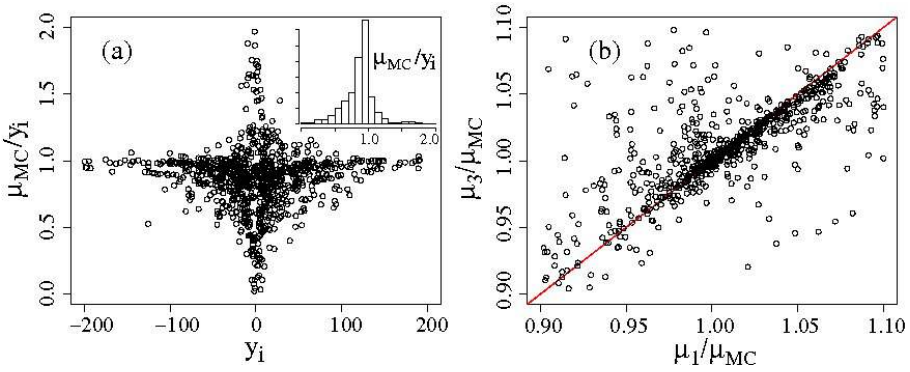

FIG 7. *Variational calculation of marginal means. Parameters as in Fig. 6 with* $n = 5$. *(a) Marginal mean* μ_{MC} *calculated with MCMC as a function of the corresponding observation* y_i. *(b) Ratio* μ_3/μ_{MC} *as a function of* μ_1/μ_{MC} *with* μ_1 *and* μ_3 *marginal means obtained from one- and three-component VB.*

Finally, Fig. 7 displays some results on the posterior means for $\alpha = \beta = 0.35$. In order to provide an additional perspective on the model, Fig. 7(a) shows the ratio μ_{MC}/y_i of the posterior mean μ_{MC} evaluated from MCMC and the actual observation y_i as a function of y_i itself. As might be expected from the model structure, the posterior mean is very close to y_i for large $|y_i|$, but becomes more variable and typically smaller in modulus than y_i when y_i tends towards 0. The histogram shows the frequency distribution for the values of the ratio μ_{MC}/y_i.

Figure 7(b) indicates that the VB estimation of the posterior means is typically very accurate even in this case of strong non-Gaussianity. In almost 90% of all cases, μ_1/μ_{MC} lies between 0.8 and 1.2. Mixture VB does not provide an obvious benefit in this regard as μ_1 and μ_3 are mostly very close to each other. For $\alpha = \beta = 1.2$ it is found that the spread in μ_1/μ_{MC} and μ_3/μ_{MC} is even further reduced (between 0.97 and 1.03).

6. The Bayesian lasso

In sparse regression, one solves the minimization problem

$$\min_{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} L(y_i - \mu - \mathbf{x}_i\boldsymbol{\beta}) + \sum_{j=1}^{p} P(|\beta_j|, \lambda) \tag{30}$$

in order to obtain shrinkage estimates for the regression coefficients $\boldsymbol{\beta}$. In (30), $(y_1, \ldots, y_n)^T \equiv \mathbf{y}$ are observations, μ is an overall mean and $\mathbf{x}_i$ a row of the $n \times p$-dimensional design matrix $\mathbf{X}$. The loss function to be minimized is denoted L, and P is a penalty function that shrinks the estimates towards zero. The amount of shrinkage is controlled by the penalty parameter λ. The popular lasso is obtained for quadratic loss and the penalty function $P(|\beta_j|, \lambda) = \lambda|\beta_j|$ (Tibshirani 1996).

Recently, Bayesian versions of sparse regression have attracted considerable attention (see, e.g., Park and Casella (2008), Hans (2009; 2010)). In the Bayesian formulation, the loss function is used to construct the data likelihood while the penalty gives rise to a prior for $\boldsymbol{\beta}$. For certain models, for example the Bayesian lasso, it has been possible to derive efficient MCMC samplers (Park and Casella 2008, Hans 2009). However, these samplers usually rely on sophisticated methodology, and for other models it may not be obvious how to construct suitable sampling schemes. For example, Park and Casella (2008) discuss a "Huberized" lasso for which MCMC sampling of a one-to-one Bayesian translation appears difficult. In such situations, variational Bayesian inference with Gaussian trial functions might present an interesting alternative.

Consider, e.g., a Bayesian sparse regression problem with a multivariate normal data likelihood (derived from a quadratic loss L) and a prior distribution $\exp[-\sum_j P(|\beta_j|, \lambda)]$. For Gaussian trial functions, it is straightforward to compute the contribution of the data likelihood to the energy part of the KL divergence. The prior gives rise to one-dimensional integrals $\int \mathcal{N}(\beta; \mu, \sigma^2) P(|\beta|, \lambda) d\beta$ which either may be solved analytically or can be evaluated numerically using Gauss-Hermite integrals. If such an approach turns out to be feasible, it is natural to ask what improvements in accuracy are afforded by using Gaussian mixtures as trial functions.

In the following, we will discuss this question for the Bayesian lasso. This model provides a convenient example as the availability of efficient MCMC samplers makes it easy to compare the variational approximations to Monte-Carlo results. The computational gain of the variational method is not too pronounced in this case, but for other models the variational approximation should provide a clear advantage. The Bayesian lasso is defined as

$$\mathbf{y}|\mu, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n), \tag{31}$$

$$\beta_i|\sigma^2, \lambda \sim \mathit{Laplace}(0, \sqrt{\sigma^2}/\lambda), \tag{32}$$

$$\mu \sim 1. \tag{33}$$

$\mathit{Laplace}(0, \sqrt{\sigma^2}/\lambda)$ denotes the Laplace distribution with mean 0 and scale parameter $\sqrt{\sigma^2}/\lambda$, i.e., $p_L(x; 0, \sqrt{\sigma^2}/\lambda) = \lambda/(2\sigma^2) \exp(-\lambda|x|/\sqrt{\sigma^2})$. In order to compute inferences for $\boldsymbol{\beta}$, it is convenient to center $\mathbf{y}$ and standardize $\mathbf{X}$. With a flat prior, the parameter μ can then be marginalized out (Park and Casella 2008) and is ignored in the following. It is straightforward to impose priors on λ and σ^2, but in order to work out the effects of the mixture trial functions as clearly as possible, these parameters will be assumed known. Specifically, σ^2 is set equal to its estimate from standard linear regression whereas λ is treated as a variable external parameter. Details of the variational calculations are given in Appendix D.

Data from a diabetes study ($n = 442$, $p = 10$) have been used in numerical examples in a number of papers on the lasso and its Bayesian version (e.g., Efron, Hastie, Johnstone and Tibshirani (2004), Park and Casella (2008), Hans (2009)) and will also be considered in the following. More specifically, we study the behavior of the variational approximations as the penalty parameter λ is

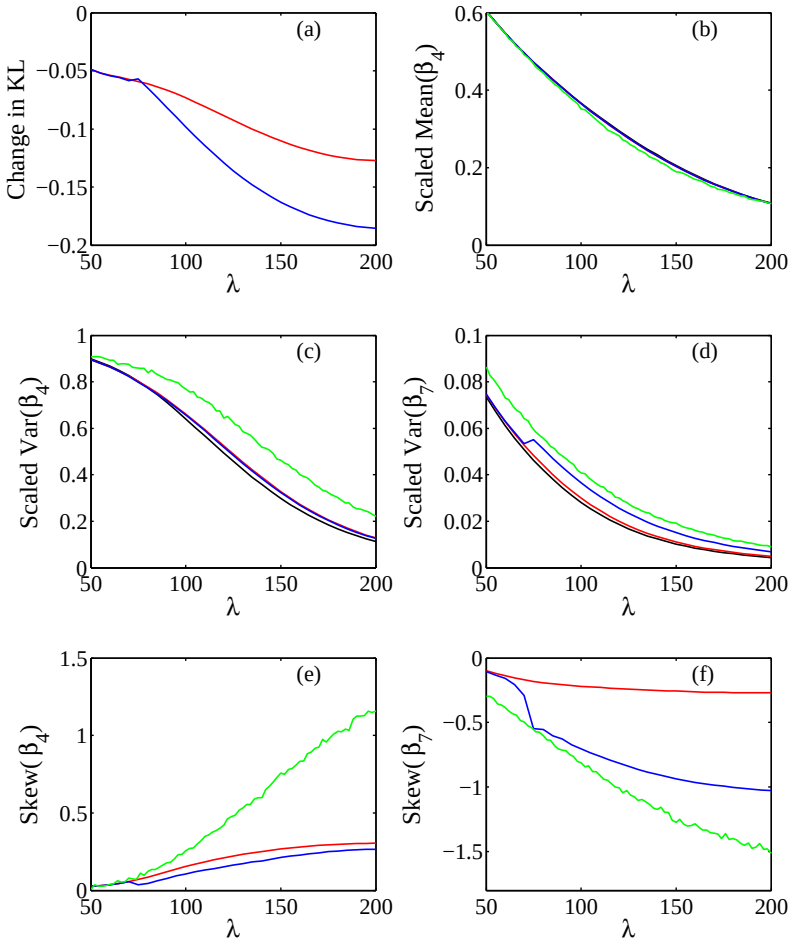

FIG 8. *Variational inference for the Bayesian lasso with diabetes data. Black, red and blue curves pertain to variational approximations with one, two, and three Gaussians, respectively, green curves are MCMC results. (a) Decrease in KL divergence. (b)-(f) Means, variances and skew for posteriors of parameters β_4 and β_7. Means and variances are scaled to values from ordinary linear regression.*

varied. Figure 8 summarizes the calculations. Variational results for a single Gaussian and two- and three-component mixtures are shown in black, red and blue, respectively, while green curves pertain to MCMC. Figure 8(a) displays the decrease in KL divergence compared to the single-Gaussian solution. Given that the maximum possible gain in KL from a two-component mixture equals $\log 2 \approx 0.69$ (Jaakkola and Jordan 1998), Fig. 8(a) indicates that even with the locking constraint two-component VB provides an appreciable and consistent improvement over the single-component approximation. The gain increases with growing λ up to values around 0.13 for $\lambda \approx 200$. The three-component solution provides further improvement mainly for larger values of λ.

Figure 8(b) shows a representative result for the mean value of the $\boldsymbol{\beta}$ posteriors, here for the example of parameter β_4 (blood pressure) normalized to its value at $\lambda = 0$. Typically, the one-component mean already provides a very good approximation to the MCMC result so that the further improvements by the mixture solutions are small.

For the variance, the difference between variational and MCMC calculations becomes larger. For some regression parameters, the mixture results show noticeable improvements, for example for the parameters β_4 and β_7 (Figs. 8(c,d)). Compared to the single-component solution, the relative increases in variance are between 10 and 50% at larger λ's. Note, in paticular, the additional improvement for the three-component solution in 8(d). For most of the other parameters, the relative increase in variance is smaller and of the order of a few percent.

For most parameters, the variational approximation of the skew correctly predicts the sign but underestimates the actual value, similar to what was found in Sec. 5. Results for β_4 and β_7 are shown in Figs. 8(e) and (f). The relatively close agreement of the three-component solution with MCMC in Fig. 8(f) is unusual. Parameters β_3 and β_9 show somewhat different behavior, there the MCMC skew stays close to 0 whereas VB predicts values around -0.1.

The KL distance for the mixture trial functions typically has more than one local minimum. In the two-component case, another solution was found[5] which has a smaller KL divergence at smaller λ, i.e., at these values it provides a better overall approximation. Compared to the single-component result, it improves the moments of the posterior marginals mainly for β_8 with the other parameters remaining almost unchanged. As the solutions of Fig. 8 mainly enhance β_4 and β_7, these observations suggest that the mixture approximations tend to focus their impact on a subset of parameters. This impression was confirmed in calculations with other data sets where mixtures typically increased variance estimates by up to 10% for most parameters, but significantly larger increases were observed for some parameters.

For the calculations of Fig. 8, the computation time for two- and three-component optimizations was increased by factors of about 3.3 and 10.4 on average compared to one-component runs (which took about 1.2 seconds on average). However, since the most time-consuming contribution comes from calculating the KL distance these factors can be expected to vary from problem to problem. For the lasso, the energy computations were almost instantaneous compared to the mixture entropy (the hypergeometric function was linearly interpolated from a dense precomputed grid). In other cases, the energy computations may be much more time-consuming thus modifying the time ratios. Another aspect that might warrant further exploration are the convergence criteria of the VB optimization routine. In particular for mixtures, it often appeared that before stopping many iterations were performed with very little change in the KL divergence. Computing 100,000 iterations of the MCMC sampler of Park and Casella (2008) in the same computational environment took about 12 seconds.

[5]The discontinuity in the three-component results in Fig.8 (blue curves) around $\lambda = 70$ is also due to a switch between different families of solutions.

7. Summary and conclusions

An important objective of current research on Variational Bayesian methods is to develop and investigate improved approximation schemes. In this context, the purpose of the present paper is to study the use of Gaussian mixture distributions as trial functions. The main results can be summarized as follows.

(i) The key problem regarding the use of mixtures as trial functions is the calculation of the entropy which becomes analytically intractable. In the present work, this obstacle was overcome with the help of a lower-bound approximation to the entropy in terms of all contributions from individual components and their pairwise combinations. (ii) Even for Gaussian mixtures, the pairwise combinations still cannot be calculated analytically. In order to make practical calculations feasible, "shape- and orientation-locking" constraints were imposed on the Gaussian covariance matrices so that the entropy approximation could be computed by standard two-dimensional numerical integration. (iii) Two simple sufficient criteria were derived that permit to check if a mixture approximation is guaranteed to provide an improvement to single-component VB. (iv) To illustrate the method, three examples were discussed in detail with target distributions chosen as isotropic generalized normal distributions, a non-Gaussian state space model and the Bayesian lasso, respectively. These examples suggest that for heavy-tailed distributions, Gaussian-mixture VB can be expected to indeed lead to an improved approximation of the target distribution in the sense of a reduced KL divergence. Appreciable improvements for posterior variance estimates could be obtained already with two- and three-component mixtures. In addition, it was shown that mixture VB provides an estimate of skewness, which is impossible for single-component Gaussian VB.

A major objective for further methodological work on the mixture approach is the development of more flexible ways to calculate the mixture entropy that allow us to overcome the current restriction to "shape- and orientation-locked" Gaussians. The calculation of Gaussian mixture entropies is a topic of ongoing research to which no final solutions have been found yet. However, in the present context the problem is alleviated to some degree by the fact that only the entropy of two-component mixtures is required, in contrast to the general k-component problem. This should make it easier to find more general numerical or even analytical approximations.

The numerical examples suggest a potential limitation of the approach in the form of a reduced effectiveness for higher-dimensional target distirbutions. This behavior might at least in part be due to the restrictions on the Gaussian covariances imposed by the locking condition; however, it is also found for the isotropic target of Sec. 3.2 where these restrictions should be less of an issue. A possible qualitative explanation might be given by the reduction in the "entropic repulsion" discussed in Sec. 3.1. Further work and investigation of additional examples is necessary to clarify this issue. However, it should be kept in mind that VB can be of significant practical value even in lower-dimensional problems. Due to the *relative* speed advantage compared to MCMC, considerable gains in *absolute* computation time can be obtained if a larger number of inference prob-

lems has to be solved. An example for such a situation is given by the inferences for multipe values of the penalty parameter in sparse regression, see Sec. 6. Another interesting lower-dimensional application of the present approach might be found in semiparametric VB approximations when the Gaussian factor remains of lower dimension.

Regarding the practical aspects of mixture VB calculations, it should be emphasized that the mixture entropy calculations need to be coded only once in a problem-independent way, and the method then requires no further essential adaptations to specific targets, except for what is already needed for the single-component calculations. It is hoped that the modest implementation effort and the potential gain in accuracy outweigh the increase in computation time and make the method an interesting, practically viable option for improving VB calculations.

Acknowledgements

Valuable comments and suggestions by two anonymous referees and the Associate Editor are gratefully acknowledged. This work was supported by an EPSRC Statistics Mobility fellowship.

Appendix A: Numerical calculation of pairwise entropy

In the following, we outline the numerical calculation of the pairwise entropy contribution S_2 defined in (4) for multivariate normal distributions under the locking condition. Essentially, one needs the integral

$$I(\mu_1, \mu_2, \Sigma_1, \Sigma_2, w_1, w_2) = \int dx_{1:k}\, \mathcal{N}(x; \mu_1, \Sigma_1) \ln\left(1 + \frac{w_2 \mathcal{N}(x; \mu_2, \Sigma_2)}{w_1 \mathcal{N}(x; \mu_1, \Sigma_1)}\right). \tag{A.1}$$

We now define the matrix U as square root of Σ_1^{-1}, i.e., $\Sigma_1^{-1} = U^T U$ with U^T the transpose of U, and set $\mu := \mu_1 - \mu_2$. Using these definitions, the above integral can be transformed to the "normal form"

$$I(\mu_1, \mu_2, \Sigma_1, \Sigma_2, w_1, w_2) = \int dx_{1:k}\, \mathcal{N}(x; 0, \mathrm{I}) \ln\left[1 + \frac{w_2 \mathcal{N}(x; U\mu, U\Sigma_2 U^T)}{w_1 \mathcal{N}(x; 0, \mathrm{I})}\right] \tag{A.2}$$

with I the identity matrix. We now invoke the locking condition $\Sigma_2 = \lambda^2 \Sigma_1$ which simplifies the above expression to

$$I(\mu_1, \mu_2, \Sigma_1, \Sigma_2, w_1, w_2) = \int dx_{1:k}\, \mathcal{N}(x; 0, \mathrm{I}) \ln\left[1 + \frac{w_2 \mathcal{N}(x; U\mu, \lambda^2 \mathrm{I})}{w_1 \mathcal{N}(x; 0, \mathrm{I})}\right]. \tag{A.3}$$

The integral can now be evaluated in cylindrical coordinates (ρ, z) with the cylinder axis directed along $U\mu$. This yields

$$I = \int_{-\infty}^{\infty} dz \int_0^{\infty} d\rho\, \Omega_{k-1} \rho^{k-2} \frac{1}{\sqrt{(2\pi)^k}} \exp\left[-\frac{1}{2}(z^2+\rho^2)\right] \times \ln\left[1 + \frac{w_2}{w_1 \lambda^k} \exp\left(-\frac{1}{2\lambda^2}(z-r)^2 + \frac{1}{2}z^2 - \frac{1}{2}(\lambda^{-2}-1)\rho^2\right)\right] \quad \text{(A.4)}$$

with $\Omega_{k-1} = 2\pi^{(k-1)/2}/\Gamma((k-1)/2)$ the surface area of a $(k-1)$-dimensional unit sphere. Furthermore, $r = |U\mu|$ which is readily evaluated from $r^2 = \mu^T \Sigma_1^{-1} \mu$. Due to the presence of the Gaussian weight function, the integral (A.4) can very efficiently be computed using generalized Gauss-Hermite quadrature. In typical calculations, we use 20 grid points for z and 10 points for (positive) ρ. To save computation time, only (z, ρ) grid points are considered for which the product of the quadrature weights is larger than a cutoff of 10^{-8}. One might also consider schemes with variable accuracy which is increased once the calculation approaches convergence. For conjugate gradient minimization, one also requires the partial derivatives of I with respect to λ, w_i, and r. These are readily obtained by differentiating the integrand in (A.4).

Equation (A.4) shows that the integral I depends only on the parameters r, λ and the ratio $\omega = w_2/w_1$, i.e., $I = I(\lambda, \omega, r)$. The total negative pairwise entropy $\bar{S}_2 = -S_2[w_1 q_1, w_2 q_2]$ is then given by

$$\bar{S}_2[w_1 q_1, w_2 q_2] = w_1 I(\lambda, \omega, r) + w_2 I(\lambda^{-1}, \omega^{-1}, r/\lambda). \quad \text{(A.5)}$$

Appendix B: Proofs of Theorems and Corollaries in Section 4

Proof of Theorem 1. The functions in $\mathcal{Q}_0^{(k)}$ are given by $\sum_i w_i q(\theta^{(i)} = \theta_0)$ with $\sum_i w_i = 1$, $w_i \geq 0$. For these functions

$$\begin{aligned} \partial_{\theta^{(j)}} \mathrm{KL}^{(k)} &= \int w_j \partial_\theta q(\theta_0) \log\left[\sum_i w_i q(\theta_0)\right] \\ &\quad + \int \sum_i w_i q(\theta_0) \frac{1}{\sum_i w_i q(\theta_0)} w_j \partial_\theta q(\theta_0) - \int w_j \partial_\theta q(\theta_0) \log p \\ &= w_j \left[\int \partial_\theta q(\theta_0) \log q(\theta_0) - \int \partial_\theta q(\theta_0) \log p\right] = 0, \end{aligned}$$

the bracket in the last line vanishing because θ_0 is a stationary point of the single-component KL distance. As $\mathrm{KL}^{(k)}$ does not depend on w_i for the functions in $\mathcal{Q}_0^{(k)}$, we can immediately conclude that $\partial_{w_j} \mathrm{KL}^{(k)} = 0$. □

Proof of Theorem 2. Straightforward calculation shows that

$$\begin{aligned} \partial_{\theta^{(j)}, \theta^{(l)}} \mathrm{KL}^{(k)} &= \mathrm{diag}(w_1, \ldots, w_k) \otimes A + \left(w w^T\right) \otimes B \\ &= \mathrm{diag}(w_1, \ldots, w_k) \otimes A \\ &\quad + \left(w_1 B^{1/2}, \ldots, w_k B^{1/2}\right)^T \left(w_1 B^{1/2}, \ldots, w_k B^{1/2}\right), \end{aligned} \quad \text{(B.1)}$$

with w a column vector containing the weights w_i, $B = B^{1/2}B^{1/2}$ and $\otimes$ the Kronecker matrix product. We can now make use of the general matrix identity $\det(M + Q^T Q) = \det\left(1_s + QM^{-1}Q^T\right) \det M$ for $r \times r$-dimensional, invertible M and $s \times r$-dimensional Q with arbitrary r and s. Assuming for now that A and B are invertible, we can apply this identity to (B.1) to obtain

$$\begin{aligned}
\det\left(\partial_{\theta^{(j)},\theta^{(l)}} \mathrm{KL}^{(k)}\right) &= \det\left(1_p + B^{1/2}A^{-1}B^{1/2}\sum_{i=1}^{k} w_i\right)(\det A)^k \prod_{i=1}^{k} w_i \\
&= \det\left[B^{1/2}\left(B^{-1} + A^{-1}\right)B^{1/2}\right](\det A)^k \prod_{i=1}^{k} w_i \\
&= \det\left[B\left(B^{-1} + A^{-1}\right)A\right](\det A)^{k-1} \prod_{i=1}^{k} w_i \\
&= (\det A)^{k-1} \det(A + B) \prod_{i=1}^{k} w_i \qquad \text{(B.2)}
\end{aligned}$$

which proves (23) for invertible A and B. However, due to continuity (23) also holds in case A or B are not invertible. To show (24), we need to study the characteristic polynomial $\det(\partial_{\theta^{(j)},\theta^{(l)}} \mathrm{KL}^{(k)} - \lambda 1_{kp})$ at $w_i = 1/k$. However, brief inspection of (B.1) reveals that we can evaluate this determinant in the same way as above if we replace A by $A - k1_p$ in (B.1). This immediately yields (24). □

Proof of Corollary 1. The k-component mixture with $(\theta^{(1)} = \theta_0, \ldots, \theta^{(k)} = \theta_0)$ and $w_1 = \cdots = w_k = 1/k$ describes the same distribution as $q(\theta_0)$, its KL divergence equalling $\mathrm{KL}^{(1)}(\theta_0)$. However, from (24) and the assumption on A it follows that its stability matrix has negative eigenvalues, i.e., the KL divergence $\mathrm{KL}^{(k)}$ has a saddle point at this distribution. Hence, the global minimum of $\mathrm{KL}^{(k)}$ must lie below the minimum of $\mathrm{KL}^{(1)}(\theta_0)$. □

Proof of Corollary 2. Consider first the case $k = 2$. From (24) follows that the stability matrix for the mixture distribution with $(\theta^{(1)} = \theta_0, \ldots, \theta^{(k)} = \theta_0)$ and $w_1 = \cdots = w_k = 1/k$ has only positive eigenvalues. By the implicit function theorem, the eigenvalues of the stability matrix are continuous functions of the weights w_1 and w_2. As long as w_1 and w_2 are both different from zero, (23) thus guarantees that the eigenvalues will remain positive, as the determinant cannot become zero. If $w_1 = 0$ or $w_2 = 0$, we have a one-component distribution which is already known to minimize the KL distance under variation of θ. For $k > 2$, we can prove the corollary inductively. As long as all mixture weights are different from zero, the positivity of the eigenvalues of the stability matrix follows in the same way as in the case of $k = 2$. If one or more weights are zero, the distribution corresponds to a mixture with a number of components less than k, and the positivity of the eigenvalues follows from the induction hypothesis. □

Appendix C: Variational calculations for non-Gaussian state-space model

To compute the energy term of the KL divergence we need the integral

$$\begin{aligned} & \int \mathcal{N}(x_{1:n};\mu,\Sigma)\ln p(x_{1:n}|y_{1:n})dx_{1:n} \\ = & \int \mathcal{N}(x_{1:n};\mu,\Sigma)\ln p(x_{1:n},y_{1:n})dx_{1:n} - \ln p(y_{1:n}) \end{aligned} \tag{C.1}$$

with $\mathcal{N}(x_{1:n};\mu,\Sigma)$ the n-dimensional normal distribution with mean μ and covariance matrix Σ. As the evidence $p(y_{1:n})$ is not known analytically, we can only compute the first term on the right-hand side of (C.1). Therefore, we cannot evaluate the actual KL divergence; rather, the optimization problem will eventually provide a lower bound on the evidence. Using (28), we obtain from (C.1)

$$\begin{aligned} & \int \mathcal{N}(x_{1:n};\mu,\Sigma)\ln p(x_{1:n},y_{1:n})dx_{1:n} \\ = & \int \mathcal{N}(x_{1:n};\mu,\Sigma)\sum_{i=1}^{n}\left[\ln p_{gn}(x_i;\phi x_{i-1},\beta,\kappa)+\ln p_{gn}(y_i;x_i,\alpha,\rho)\right]dx_{1:n} \\ = & \; n\ln\left[\frac{\beta}{2\kappa\Gamma(1/\beta)}\frac{\alpha}{2\rho\Gamma(1/\alpha)}\right] \\ & -\kappa^{-\beta}\int \mathcal{N}(x_1;\mu_1,\Sigma_{11})|x_1|^{\beta}dx_1-\rho^{-\alpha}\sum_{i=1}^{n}\int \mathcal{N}(x_i;\mu_i,\Sigma_{ii})|y_i-x_i|^{\alpha}dx_i \\ & -\kappa^{-\beta}\sum_{i=2}^{n}\int \mathcal{N}(x_{i-1:i};\mu_{i-1:i},\Sigma_{i-1:i})|x_i-\phi\,x_{i-1}|^{\beta}dx_{i-1:i} \end{aligned} \tag{C.2}$$

where the normal distributions for x_i and $x_{i-1:i}$ in the last two lines refer to the corresponding marginals of $\mathcal{N}(x_{1:n},\mu,\Sigma)$. The two one-dimensional integrals in the second-to-last line of (C.2) can be solved analytically in terms of confluent hypergeometric functions using the relation

$$\int \mathcal{N}(x,\mu,\sigma^2)|x|^{\gamma}dx=\sqrt{\frac{(2\sigma^2)^{\gamma}}{\pi}}\,\Gamma\left(\frac{1+\gamma}{2}\right)\,{}_1F_1\left(-\frac{\gamma}{2},\frac{1}{2},-\frac{\mu_i^2}{2\sigma^2}\right) \tag{C.3}$$

(obviously, as a function of μ the right-hand side simply provides a smoothed version of $|\mu|^{\gamma}$). For the computation of the confluent hypergeometric function ${}_1F_1$, efficient numerical routines are available (Zhang and Jin 1996). To evaluate the double integral of (C.2) involving x_{i-1} and x_i, we first integrate out one variable analytically using (C.3) and then perform the second integration using Gauss-Hermite quadrature.

Equation (C.2) shows that the energy only depends on the elements $\Sigma_{i,i}$ and $\Sigma_{i\pm1,i}$ of the covariance matrix Σ. For VB with a single Gaussian, (11)

thus implies that the inverse of the optimized covariance matrix will be tridiagonal. The single-Gaussian VB problem therefore involves the optimization of an n-component mean vector μ and a covariance matrix Σ depending on $2n-1$ parameters. For VB with mixtures, there is no equivalent to relation (11) that would impose a restriction on the form of the optimized covariance matrix. However, the numerical calculations show that, at least under the locking condition, the optimized inverse covariance matrices retain the band structure form to a very good degree of approximation. It is therefore reasonable to impose this constraint from the outset to simplify calculations. In this way, VB with two components results in a $4n+1$-dimensional optimization problem ($2n$ parameters for the two means, $2n-1$ parameters for the covariance matrix, and one parameter each for the locking ratio λ and the mixture weights). Similarly, three-component VB leads to a $5n+3$-dimensional problem. All minimizations can be performed using standard numerical methods.

Appendix D: Variational calculations for Bayesian lasso

To compute the energy term of the KL divergence we need the logarithm of the target distribution which reads

$$\log p \sim -\frac{(\tilde{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})^\top(\tilde{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} - \lambda\sum_{i=1}^{p}\frac{|\beta_i|}{\sqrt{\sigma^2}} \tag{D.1}$$

after eliminating irrelevant constants. For a trial function $\mathcal{N}(\boldsymbol{\beta};\boldsymbol{\mu},\boldsymbol{\Sigma})$ the energy term is given by

$$\begin{aligned} E(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\mu})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}) + \sum_{i,j}\boldsymbol{\Sigma}_{ij}(\mathbf{X}^\top\mathbf{X})_{ij}}{2\sigma^2} \\ &\quad + \frac{\lambda}{\sqrt{\sigma^2}}\sum_{i=1}^{p}\sqrt{\frac{2\Sigma_{ii}}{\pi}}\,{}_1F_1\left(-\frac{1}{2},\frac{1}{2},-\frac{\mu_i^2}{2\Sigma_{ii}}\right) \end{aligned} \tag{D.2}$$

with ${}_1F_1$ the confluent hypergeometric function (see (C.3)).

For the variational minimization, the Cholesky decomposition of $\boldsymbol{\Sigma}$ is used. With single Gaussians as trial functions, the search space for the optimization thus has dimension $(p^2+3p)/2$ whereas for mixtures of two and three Gaussians the dimensionalities are $(p^2+5p+4)/2$ and $(p^2+7p+8)/2$, respectively.

For the optimization, a number of derivative-free algorithms were tried out, and Powell's method (Press, Teukolsky, Vetterling and Flannery 2007) was found to provide the best results in terms of speed and reliability. It remains open whether performance could be further improved with the help of gradient-based algorithms. In the calculations with variational mixture approximations, the initial values for the optimization were chosen as the variational solution for the previous λ when going through a sequence of λ's.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer, New York. MR2247587

Chen, J.-Y., Hershey, J. R., Olsen, P. A., and Yashchin, E. (2008). Accelerated Monte Carlo for Kullback-Leibler divergence between Gaussian mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4553–4556.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* **32**, 407–499. MR2060166

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Appl. Statist.* **21**, 455–472.

Goldberger, J., Gordon, S., and Greenspan, H. (2003). An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)*, 487–493.

Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–845. MR2564494

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing* **20**, 221–229. MR2610774

Hershey, J. R. and Olsen, P. A. (2007). Approximating the Kullback-Leibler divergence between Gaussian mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2007*, IV-317–IV-320.

Huber, M. F., Bailey, T., Durrant-Whyte, H., and Hanebeck, U. D. (2008). On entropy approximation for Gaussian mixture random vectors. In *Proceedings of the EEE International Conference on Multisensor Fusion and Integration for Intelligent Systems 2008*, 181–188.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, ed. M. I. Jordan, MIT Press, Cambridge, MA, 163–174.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms.* Cambridge University Press, New York. MR2012999

Minka, T. (2005). Divergence measures and message passing. *Microsoft Technical Report* MSR-TR-2005-173.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation* **21**, 786–792. MR2478318

Opper, M. and Saad, D. (eds.) (2001). *Advanced mean field methods: theory and practice.* Neural Information Processing Series. MIT Press, Cambridge, MA. MR1863214

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. The American Statistician **64**, 140–153. MR2757005

Park, T. and Casella, G.(2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686. MR2524001

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P.(2007). *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge. MR2371990

Smidl, V. and Quinn, A. (2006). *The Variational Bayes Method in Signal Processing.* Springer, Berlin, Heidelberg, New York.

Tibshirani, R.(1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–288. MR1379242

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learning* **1**, 1–305.

Zhang, S. and Jin, J. (1996). *Computation of Special Functions.* Wiley, New York. MR1406797

Zobay, O. (2009). Mean field inference for the Dirichlet process mixture model. *Electronic Journal Statistics* **3**, 507–545. MR2519531