Electronic Journal of Statistics
Vol. 17 (2023) 1893–1922
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2142

Bayesian inference and prediction for mean-mixtures of normal distributions*

Pankaj Bhagwat and Éric Marchand

Université de Sherbrooke,
Département de mathématiques,
Sherbrooke Qc, CANADA, J1K 2R1
e-mail: pankaj.uttam.bhagwat@usherbrooke.ca; eric.marchand@usherbrooke.ca

Abstract: We study frequentist risk properties of predictive density estimators for mean mixtures of multivariate normal distributions, involving an unknown location parameter $\theta \in \mathbf{R}^d$, and which include multivariate skew normal distributions. We provide explicit representations for Bayesian posterior and predictive densities, including the benchmark minimum risk equivariant (MRE) density, which is minimax and generalized Bayes with respect to an improper uniform density for θ. For four dimensions or more, we obtain Bayesian densities that improve uniformly on the MRE density under Kullback-Leibler loss. We also provide plug-in type improvements, investigate implications for certain type of parametric restrictions on θ, and illustrate and comment the findings based on numerical evaluations.

MSC2020 subject classifications: Primary 62C20, 62C25; secondary 62C10.
Keywords and phrases: Bayes predictive density, dominance, Kullback-Leibler loss, minimax, minimum risk equivariant, mean mixtures, multivariate normal, skew-normal distribution.

Received January 2022.

1. Introduction

1.1. Motivation and scientific context

The findings of this paper relate to predictive density estimation for mean-mixture of normal distributions. The modelling of data via mixing multivariate normal distributions has found many applications and lead to methodological challenges for statistical inference. These include finite mixtures, as well as continuous mixing on the mean and/or the variance. Whereas scale or variance mixtures of multivariate normal distributions compose a quite interesting subclass of spherically symmetric distributions, modelling asymmetry requires mixing on the mean and prominent examples are generated via variance-mean mixtures (e.g., [7]), as well as mean-mixtures of multivariate normal distributions (e.g., [1, 5]) and references therein). Moreover, such mean-mixtures, which are the subject of study here, generate or are connected to multivariate skew normal distributions (e.g., [6]) which have garnered much interest over the years.

*Éric Marchand's research is supported by the Natural Sciences and Engineering Research Council of Canada. Pankaj Bhagwat is grateful to the ISM (Institut des sciences mathématiques) for financial support.

The development of shrinkage estimation techniques, namely since Stein's inadmissibility finding ([26]) concerning the maximum likelihood or best location equivariant estimator under squared error loss in three dimensions or more, has had a profound impact on statistical theory, thinking, methods, and practice. Examples include developments on sparsity and regularization methods, empirical Bayes estimation, adpative inference, small area estimation, nonparametric function estimation, and predictive density estimation. Cast in a decision-theoretic framework, Stein's original result has been expanded in many diverse ways, namely to other distributions or probability models, and namely for spherically symmetric and elliptically symmetric distributions (see for instance, [11]). There have been fewer findings for multivariate skew normal or mean-mixtures of normal distributions, but the recent work of Kubokawa et al. [17] establishes point estimation minimax improvements of the best location equivariant estimator under quadratic loss, when the dimension of the location parameter is greater than or equal to four, and with underlying known perturbation parameter.

Predictive density estimation has generated much interest over the past twenty years or so, and addresses fundamental issues in statistical predictive analysis. Decision-theoretic links between shrinkage point estimation and shrinkage predictive density estimation for normal models have surfaced (e.g., [14], [13]) and stimulated much activity (see for instance [12]), including findings for restricted parameter spaces (e.g., [10, 16]). The main objective of this work is thus to explore the problem of predictive density estimation for mean-mixtures of normal (MMN) distributions. A secondary objective is to provide novel representations for Bayesian posterior distributions and predictive densities for MMN models, which have been found to be lacking in the literature.

Following early findings of Komaki (e.g., [14]) on the predictive density estimation problem for multivariate normal models under Kullback-Leibler loss, George, Liang and Xu in [13] exhibited further parallels with the point estimation problem for normal distribution under quadratic loss. They provide sufficient conditions on marginal distributions and prior distributions to get improved shrinkage predictive density estimators when the dimension is greater than or equal to three. Thus, motivated by these connections, it is interesting to investigate whether such shrinkage plays any role in the predictive density estimation problem for mean-mixture of multivariate normal models and we focus on frequentist risk efficiency of predictive density estimators under Kullback-Leibler loss. Our contribution here consists in identifying classes of plug-in type predictive densities and of Bayes predictive densities which are minimax and dominate the benchmark minimum equivariant estimator (MRE) for the case when the dimension of the location parameter is greater than or equal to four.

1.2. The prediction problem

The prediction problem that we study relates to mean-mixtures of normal distributions, which are a large class of distributions including multivariate skew normal distributions and defined as follows.

Definition 1.1. *A random vector $X \in \mathbf{R}^d$ is said to have a mean-mixture of normal distributions, denoted as $X \sim MMN_d(\theta, a, \Sigma, \mathcal{L})$, if it admits the representation*

$$X|V = v \sim N_d(\theta + va\,, \Sigma)\,,\, V \sim \mathcal{L}\,, \tag{1.1}$$

where $\theta \in \mathbf{R}^d$ is a location parameter, $a \in \mathbf{R}^d - \{0\}$ is a known perturbation vector, Σ is a known positive definite matrix, and V is a scalar random variable with cdf $\mathcal{L}$.

Consider then $X|\theta \sim MMN_d(\theta, a, \Sigma_X, \mathcal{L}_1)$ and $Y|\theta \sim MMN_d(\theta, a, \Sigma_Y, \mathcal{L}_2)$, independently distributed and let Let $p(x|\theta)$ and $q(y|\theta)$ denote the conditional densities of X and Y given θ, respectively. Based on observing $X = x$, we consider the problem of finding a suitable predictive density estimator $\hat{q}(y; x)$ for $q(y|\theta)\,, y \in \mathbf{R}^d$. The ubiquitous Kullback-Leibler (KL) divergence between two Lebesgue densities f and g on $\mathbf{R}^m$, defined as

$$\rho(f, g) \;=\; \int_{\mathbf{R}^m} f(t)\, \log \frac{f(t)}{g(t)}\, dt\,, \tag{1.2}$$

is the basis of Kullback-Leibler loss given by

$$L(\theta, \hat{q}) \;=\; \rho(q_\theta, \hat{q})\,. \tag{1.3}$$

We evaluate the performance of the density estimators using KL loss in (1.3), and the associated KL risk function

$$R_{KL}(\theta, \hat{q}) = \int_{\mathbf{R}^d} \{\int_{\mathbf{R}^d} q(y|\theta)\, \log \frac{q(y|\theta)}{\hat{q}(y; x)}\, dy\}\, p(x|\theta)\, dx. \tag{1.4}$$

A benchmark predictive density estimator for $q(y|\theta), y \in \mathbf{R}^d$, is given by the Bayes predictive density estimator $\hat{q}_U(y; X), y \in \mathbf{R}^d$, with respect to the uniform prior density on $\mathbf{R}^d$. It is known to be the minimum risk equivariant (MRE) predictive density estimator under changes of location, as well as minimax. We will expand more on these properties below as well as analytical representations in Section 2.2, but the main problem relates to uniform improvements in terms of risk for unconstrained $\theta \in \mathbf{R}^d$ and for some restricted parameter spaces $\theta \in C \subset \mathbf{R}^d$. We provide predictive densities that dominate $\hat{q}_U(\cdot; X)$ among plug-in type densities and Bayesian densities. A secondary problem of interest which we address is that of making available exact and useful expressions for Bayes posterior and predictive analyses.

1.3. Summary of main findings

For the problem described above, here is a short summary of the main results of the paper.

- **MRE density:** We provide the general representation $\hat{q}_U \sim MMN_d(X, a,\; \Sigma_X + \Sigma_Y, \mathcal{L}_3)$; with $\mathcal{L}_3$ defined in Theorem 2.1; of the minimum risk equivariant (MRE) density estimator as well as several examples.

- **Bayesian analysis under Gaussian priors:** We derive novel and appealing mixture representations for Bayesian posterior and predictive densities under Gaussian priors on θ.
- **Domination results (Plug-in type improvements):** We provide for $d \geq 4$ sufficient conditions for a plug-in type density of the form $MMN_d(\hat{\hat{\theta}}(X), a, \Sigma_X + \Sigma_Y, \mathcal{L}_3)$ to dominate the MRE density (equivalently to be minimax), and which relate to dominance conditions for an estimator of the mean of a $d-1$ dimensional normal distribution to be minimax under squared error loss.
- **Domination results (Improved Bayes predictive densities):** We obtain Bayes predictive densities that dominate the MRE density for $d \geq 4$, (equivalently that are minimax), which relate to Bayesian shrinkage estimators of a multivariate normal mean in $d-1$ dimensions under squared error loss.
- **Restricted parameter spaces:** We also extend the dominance results to cases involving certain types of parametric restrictions and which are also applicable for $d = 2, 3$.

The organization of this manuscript is as follows. Section 2.1 contains several introductory definitions, properties and examples of MMN models, including a useful canonical form which subdivides the MMN distributed vector into d independent components, one of which a univariate MMN distribution and the others as normal distributions. Section 2.2 focuses on the MRE density with a useful representation accompanied by various examples. Section 2.3 expands on the calculation of minimax risk and a representation in terms of the entropy of a univariate distribution. Section 3 is devoted to Bayesian posterior and predictive analysis with several novel representations. Sections 4.1 and 4.2, namely Theorems 4.1, Theorem 4.2 and Corollary 4.2, contain the main dominance results, with plug-in type and Bayesian improvements. In both cases, the main technique employed rests upon the canonical transformation presented in Section 2.1 and permits to split up the KL risk as the addition of two parts, one of which can be operated on using known normal model prediction analysis findings. Section 4.3 deals with parametric restrictions and further applications of Theorems 4.1 and 4.2. Finally in Section 5, some further illustrations are provided and accompanied by numerical frequentist risk comparisons.

2. Preliminary results and definitions

Here are some details and properties of mean-mixture of normal distributions, its canonical form, the MRE predictive density, and minimax risk. In the following, we will denote $\phi_d(z; \Sigma)$ the probability density function of a $N_d(0, \Sigma)$ distribution evaluated at $z \in \mathbf{R}^d$ and for positive definite Σ. When $\Sigma = I_d$, we may simplify the writing to $\phi_d(z)$, and then for $d = 1$ to $\phi(z)$. We will denote Φ as the cdf of a $N(0,1)$ distribution. The organization of the following subsections is as follows.

In Section 2.1, we extract some stochastic properties of MMN distributions, and link such distributions to multivariate skew normal distributions. Lemma 2.1 describes a key orthogonal transformation crucial to the main dominance findings of Section 4, which splits up a d-dimensional MMN distributed vector into d independent univariate components, $d-1$ of which are normally distributed and the other MMN distributed. As a complement to Section 2.1, we provide in the Appendix a result (Lemma 6.1) which facilitates the representation of various MMN densities.

In Section 2.2, as a continuation to Section 1.2 where the prediction problem is cast into a decision-theoretic framework with the specification of Kullback-Leibler loss and risk, we focus on the benchmark best equivariant predictive density which admits an appealing representation and which is minimax (Theorem 2.1). We conclude the section with several examples varying with the choice of model for the observable X and predicted Y.

In Section 2.3, we provide a decomposition of KL risk and obtain a simplification of the constant and minimax risk of the MRE predictive density. This brings into play interesting expressions on their own relative to the entropy of MMN distributions.

2.1. The model

As mentioned above, the distributions of interest are mean-mixtures of multivariate normal (MMN) distributions, both for our observables and densities to be estimated by a predictive density estimator. Such distributions connect to multivariate skew normal distributions and have been the object of interest in recent work with studies of stochastic properties (e.g., [1], [5], [25]), and shrinkage estimation about its location parameter ([17]).

As an alternative and equivalently to Definition 1.1, the random vector X has stochastic representation

$$X = \theta + \Sigma^{1/2} Z + V a\,, \tag{2.1}$$

where $Z \sim N_d(0, I_d)$ and $V \sim \mathcal{L}$ on $\mathbf{R}$, and its probability density function can be expressed as:

$$\begin{aligned} p(x|\theta) &= \mathbf{E}^V \left\{\phi_d\left(x-\theta-Va, \Sigma\right)\right\} \\ &= \phi_d\left(x-\theta, \Sigma\right)\, \mathbf{E}^V \left(e^{-\frac{V^2}{2} a^T \Sigma^{-1} a} e^{V\,(x-\theta)^T \Sigma^{-1} a}\right)\,. \end{aligned} \tag{2.2}$$

Thus, we note that the density function of a MMN random vector can be decomposed in two parts: one symmetrical density $\phi_d(\cdot)$, and the other part a function of the projection of $(x-\theta)$ in the direction of $\Sigma^{-1}a$. Moreover, this construction isolates the asymmetry in the direction $\Sigma^{-1}a$ and the scale is controlled by the random variable V.

Remark 2.1. *It is easy to see that the family of MMN distributions is closed under linear combinations of independent components. Specifically, if* $X_i|\theta \sim MMN_d(\theta, a, \Sigma_i, \mathcal{L}_i)$*,* $i = 1, \ldots, n$*, are independently distributed, then* $\sum_{i=1}^{n} b_i X_i|\theta$

$\sim MMN_d((\sum_{i=1}^n b_i)\,\theta, a, \sum_{i=1}^n b_i^2\,\Sigma_i, \mathcal{L}_0)$ with $\mathcal{L}_0$ the cdf of the mixing variable $V_0 =^d \sum_{i=1}^n b_i V_i$. Namely, for the identically distributed case with $\Sigma_i = \Sigma$ and the sample mean with $b_i = 1/n$, we obtain that

$$\bar{X}|\theta \sim MMN_d(\theta, a, \Sigma/n, \mathcal{L}_0)\,, \text{ with } \mathcal{L}_0 \text{ the cdf of } \bar{V}\,.$$

It thus follows, as observed in [17], that findings applicable for a single MMN distributed observable X can be extended to the random sample case.

We now turn our attention to a fundamental and useful decomposition, or canonical form, ([1]) for MMN distributions which follows rather directly from Definition 1.1.

Lemma 2.1. *For a random vector $X \sim MMN_d(\theta, a, \Sigma, \mathcal{L})$ as in (1.1), there exists an orthogonal matrix H such that the first row of H is proportional to $a^\top \Sigma^{-1/2}$ and $Z = H\Sigma^{-1/2}X$ has a $MMN_d(H\Sigma^{-1/2}\theta, a_0, I_d, \mathcal{L})$ distribution with $a_0 = (\sqrt{a^T\Sigma^{-1}a}, 0, \ldots, 0)^T$.*

Such a Z may be referred to as a canonical form and is comprised of d independent components. Moreover $Z - H\Sigma^{-1/2}\theta$ has $d-1$ components which are $N(0,1)$ distributed and another distributed as $MMN_1(0, a_0, 1, \mathcal{L})$. Such a canonical form construction is not unique and depends on the choice of H.

As already mentioned, the family of MMN distributions contains many interesting distributions and we expand here and in the Appendix with examples, which will also inform us for our predictive density problem and related Bayesian posterior analysis. A prominent example is the multivariate skew normal distribution due to Azzalini and Dalla Valle [6]. Indeed, if we consider $V \sim TN(0,1)$, the standard truncated normal distribution on R_+ in (1.1), we get the multivariate skew-normal family of distributions with densities

$$p(x|\theta) = 2\phi_d\left(x - \theta; \Sigma + aa^T\right)\Phi\left(\frac{(x-\theta)^\top \Sigma^{-1} a}{\sqrt{1 + a^\top \Sigma^{-1} a}}\right). \tag{2.3}$$

Here, we note that also $V \sim \sqrt{\chi_1^2}$, i.e. the square root of a Chi-square distribution with $k = 1$ degrees of freedom. Various other choices of the mixing density have appeared in the literature (e.g., [5]), namely cases where $V \sim \sqrt{\chi_k^2}$ or V is Gamma distributed. A general representation with accompanying examples are postponed to the Appendix.

2.2. The minimum risk equivariant predictive density

Consider $X|\theta \sim MMN_d(\theta, a, \Sigma_X, \mathcal{L}_1)$ and $Y|\theta \sim MMN_d(\theta, a, \Sigma_Y, \mathcal{L}_2)$, independently distributed as in Definition 1.1, i.e.

$$X|\theta, V_1 \sim N_d(\theta + V_1\, a, \Sigma_X)\,,\; Y|\theta, V_2 \sim N_d(\theta + V_2\, a, \Sigma_Y)\,, \text{with } V_1 \sim \mathcal{L}_1\,,\; V_2 \sim \mathcal{L}_2\,. \tag{2.4}$$

Bayesian predictive densities play a central role and they will be studied throughout this paper. For a prior density π for θ and a generated posterior density $\pi(\cdot|x)$ with respect to a σ-finite measure ν, it is known (e.g., [3, 4]) that the Bayes predictive density is given by

$$\hat{q}_\pi(y;x) = \int_{\mathbf{R}^d} q(y|\theta)\, \pi(\theta|x)\, d\nu(\theta)\,. \tag{2.5}$$

As mentioned in Section 1.2, the benchmark MRE predictive density $\hat{q}_U(\cdot;X)$ coincides with the Bayes predictive density estimator with respect to the uniform prior density on $\mathbf{R}^d$. In [15], a representation, which applies to both integrated squared-error loss and KL loss, for $\hat{q}_U$ is provided. For our prediction problem, the following result makes use of this representation and summarizes the above optimality properties.

Theorem 2.1. *The MRE predictive density estimator of the density of Y relative to model (2.4) under KL loss, is given by the Bayes predictive density $\hat{q}_U$ under prior $\pi_U(\theta) = 1$. Furthermore, we have*

$$\hat{q}_U(\cdot;X) \sim MMN_d(X, a, \Sigma_X + \Sigma_Y, \mathcal{L}_3)\,, \tag{2.6}$$

where $\mathcal{L}_3$ is the cdf of $V_3 = V_2 - V_1$. Finally, $\hat{q}_U(\cdot;X)$ is minimax under KL loss.

Proof. The MRE and minimax properties are given in Proposition 2 and Theorem 1 of [20], respectively. For a location family prediction problem with $X \sim p(x-\theta)$ and $Y \sim q(y-\theta)$ independently distributed, it is shown in [15] that

$$\hat{q}_U(y;X) \;=\; q * \bar{p}(y-x)\,, \text{ with } \bar{p}(t) = p(-t)\,,$$

i.e., the convolution of q and the additive inverse of p followed by a change of location equal to x. For model (1.1), the above convolution $q * \bar{p}$ is given by the density of $Y - X$ in model (1.1) with $\theta = 0$, and the result follows since

$$Y - X|V_1, V_2 \sim N_d((V_2 - V_1)\, a, \Sigma_X + \Sigma_Y). \qquad \square$$

Theorem 2.1 is quite general and can be viewed as an extension of the multivariate normal case with $a = 0$ and $\hat{q}_U(\cdot;X) \sim \mathrm{N}_d(X, \Sigma_X + \Sigma_Y)$. It is particularly interesting that the MRE density estimator also belongs to the class of MMN distributions with location parameter x and the same perturbation parameter a. As well, the distribution of the difference $V_2 - V_1$ plays a key role in Theorem 2.1's representation of the MRE predictive density, and as illustrated in the next series of examples.

Example 2.1. *When continuous, the mixing distributions can be taken to have a scale parameter equal to one without loss of generality, since a multiple can be integrated into the shape vector a. For examples (C) and (D) below, we will make use of the following lemma whose proof is relegated to the Appendix.*

Lemma 2.2. *For all $B, c \in \mathbf{R}$, $A \in \mathbf{R}_+$, we have*

$$\int_0^\infty \Phi(ct)\, e^{-\frac{t^2}{2A} + Bt}\, dt \;=\; e^{\frac{AB^2}{2}}\, \sqrt{2\pi A}\; \Phi_2\Big(\frac{cAB}{\sqrt{1+c^2A}}, B\sqrt{A}; \frac{c\sqrt{A}}{\sqrt{1+c^2A}}\Big)\,, \tag{2.7}$$

where $\Phi_2(z_1,z_2;\rho)$ the cdf evaluated at $z_1,z_2 \in \mathbf{R}$ of a bivariate normal distributions with means equal to 0, variances equal to 1 and covariance equal to ρ.

(A) *For the case of degenerate V_2 with $\mathbf{P}(V_2=v_2)=1$, i.e., when the distribution of $Y|\theta$ is normal with $Y \sim N_d(\theta+av_2,\Sigma_Y)$, the MRE predictive density reduces to $\hat{q}_U(\cdot;X) \sim MMN_d(X+av_2,-a,\Sigma_X+\Sigma_Y,\mathcal{L}_1)$.*

(B) *For the case of degenerate V_1 with $\mathbf{P}(V_1=v_1)=1$, i.e., when the distribution of X is normal with $X \sim N_d(\theta+av_1,\Sigma_Y)$, the MRE predictive density reduces to $\hat{q}_U(\cdot;X) \sim MMN_d(X-av_1,a,\Sigma_X+\Sigma_Y,\mathcal{L}_2)$.*

(C) *We consider in this example V_1,V_2 i.i.d. exponentially distributed with densities $f(t)=e^{-t}\,\mathbf{I}_{(0,\infty)}(t)$, as well as $\Sigma_X=\sigma_X^2\,I_d$ and $\Sigma_Y=\sigma_Y^2\,I_d$. Here the distribution of V_3 is Laplace or double-exponential with density $\frac{1}{2}\,e^{-|v_3|}$ on $\mathbf{R}$. Therefore, from Theorem 2.1, we have*

$$\begin{aligned}\hat{q}_U(y;x) &= \int_{\mathbf{R}} \frac{1}{2}\,e^{-|v_3|}\,\frac{1}{\sigma_S^d}\,\phi_d\Big(\frac{y-x-av_3}{\sigma_S}\Big)\,dv_3\,,\\ &= \phi_d\left(y-x;\sigma_S^2 I_d\right)\int_{\mathbf{R}_+} e^{-(v_3^2\frac{\|a\|^2}{2\sigma_S^2}+v_3)}\cosh\left(v_3\Big(\frac{(y-x)^\top a}{\sigma_S^2}\right)dv_3\,,\end{aligned}$$

with $\sigma_S=(\sigma_X^2+\sigma_Y^2)^{1/2}$. By making use of Lemma 2.2 with $A=\frac{\sigma_S^2}{\|a\|^2}$, $B=-1\pm\frac{(y-x)^\top a}{\sigma_S^2}$, and $c=0$, we obtain (for $a\neq 0$)

$$\begin{aligned}\hat{q}_U(y;x) &= \sqrt{\frac{\pi\sigma_S^2}{2\|a\|^2}}\,\phi_d(y-x;\sigma_S^2 I_d)\,e^{\frac{\sigma_S^2}{2\|a\|^2}+\frac{\{(y-x)^\top a\}^2}{2\sigma_S^2\|a\|^2}}\\ &\quad\times\left[\left\{e^{-\frac{(y-x)^\top a}{\|a\|^2}}\,\Phi\left(\frac{\sigma_S}{\|a\|}\Big(\frac{(y-x)^\top a}{\sigma_S^2}-1\Big)\right)\right\}\right.\\ &\quad\left.+\left\{e^{\frac{(y-x)^\top a}{\|a\|^2}}\,\Phi\left(-\frac{\sigma_S}{\|a\|}\Big(\frac{(y-x)^\top a}{\sigma_S^2}+1\Big)\right)\right\}\right].\end{aligned}$$

(D) *Consider V_1,V_2 i.i.d. truncated normal distributed $TN(0,1)$ (or equivalently as $\sqrt{\chi_1^2}$) for which X and Y are i.i.d. as multivariate skew normal as in (2.3). A straightforward calculation yields the density*

$$g_{V_3}(t) \;=\; 2\sqrt{2}\;\phi\Big(\frac{t}{\sqrt{2}}\Big)\;\Phi\Big(-\frac{|t|}{\sqrt{2}}\Big)\;\mathbf{I}_{\mathbf{R}}(t)\,,$$

for $V_3 =^d V_2-V_1$. It follows from Theorem 2.1, for $\Sigma_X=\sigma_X^2\,I_d$ and $\Sigma_Y=\sigma_Y^2\,I_d$, denoting $\sigma_S=(\sigma_X^2+\sigma_Y^2)^{1/2}$ again, that

$$\begin{aligned}\hat{q}_U(y;x) &= \int_{\mathbf{R}} 2\sqrt{2}\;\phi\Big(\frac{t}{\sqrt{2}}\Big)\;\Phi\Big(-\frac{|t|}{\sqrt{2}}\Big)\,\phi_d(y-x-at;\sigma_S^2 I_d)\,dt\,,\\ &= \frac{2}{\sqrt{\pi}}\,\phi_d(y-x;\sigma_S^2 I_d)\end{aligned}$$

$$\times \quad \int_{\mathbf{R}_+} \Phi(-\frac{t}{\sqrt{2}})\, e^{-\frac{t^2}{2}(\frac{1}{2}+\frac{a^\top a}{\sigma_S^2})} \left\{ e^{\frac{(y-x)^\top a t}{\sigma_S^2}} + e^{-\frac{(y-x)^\top a t}{\sigma_S^2}} \right\} dt\,.$$

Now, by making use of Lemma 2.2 with $c = -\frac{1}{\sqrt{2}}$, $A = \frac{2\sigma_S^2}{\sigma_S^2+2a^\top a}$, *and* $B = \pm\frac{(y-x)^\top a}{\sigma_S^2}$, *collecting terms, and setting* $f_k = \sqrt{\sigma_S^2 + k a^\top a}$ *for* $k = 1, 2$, *we obtain*

$$\begin{aligned} \hat{q}_U(y;x) &= \frac{4\sigma_S}{f_1}\, \phi_d\left(y-x; \sigma_S^2(I_d + \frac{aa^\top}{f_1^2})\right) \\ &\times \left\{ \Phi_2\left(-\frac{(y-x)^\top a}{f_1 f_2}, \frac{\sqrt{2}(y-x)^\top a}{\sigma_S f_2}; \frac{-\sigma_S}{\sqrt{2}\, f_1}\right) \right. \\ &+ \left. \Phi_2\left(\frac{(y-x)^\top a}{f_1 f_2}, -\frac{\sqrt{2}(y-x)^\top a}{\sigma_S f_2}; \frac{-\sigma_S}{\sqrt{2} f_1}\right) \right\}. \end{aligned}$$

In the evaluation above, we made use of the identities $(I - \frac{aa^\top}{f_2^2})^{-1} = I + \frac{aa^\top}{f_1^2}$ *and* $|I + \frac{aa^\top}{f_1^2}| = 1 + \frac{a^\top a}{f_1^2}$ *, which is a special case of the Sherman-Morrison formula for the matrix inversion of* $A + b_1 b_2^\top$ *with* A *being a square matrix and* b_1 *and* b_2 *vectors of the same dimension.*

2.3. Minimax risk and entropy

We will make use of Lemma 2.1's canonical form to transform a mean-mixture of normal distributions vector into two independent components and to capitalize on the corresponding simplification for KL divergence which is as follows. The following technical result summarizes this and it will be also of critical use in Section 4.

Lemma 2.3. *For* $i = 1, 2$, *let* f_i, g_i *be densities for* T_i, *and let* f *and* g *be the resulting joint densities of* $T = (T_1, T_2)^\top$ *associated with the* f_i*'s and the* g_i*'s respectively under the assumption of independence. Then, we have for the KL divergence defined in (1.2)*

$$\rho(f,g) \;=\; \rho(f_1,g_1) \;+\; \rho(f_2,g_2)\,. \tag{2.8}$$

Proof. By independence, we have

$$\begin{aligned} \rho(f,g) &= \mathbf{E}^T \left\{ \log\left(\frac{f_1(T_1) f_2(T_2)}{g_1(T_1) g_2(T_2)}\right) \right\} \\ &= \mathbf{E}^T \left\{ \log\left(\frac{f_1(T_1)}{g_1(T_1)}\right) \right\} + \mathbf{E}^T \left\{ \log\left(\frac{f_2(T_2)}{g_2(T_2)}\right) \right\}, \end{aligned}$$

which is (2.8). ☐

The Kullback-Leibler risk expressions brings into play the entropy associated with MMN distributions. Such a measure is not easily manipulated into a closed form (see for instance [9] for the study of entropy for multivariate skewed normal distributions), but it can be expressed in terms of the entropy of a univariate MMN distribution, as illustrated with the following expansion of the constant and minimax risk of the MRE density $\hat{q}_U$ in the context of model (2.4). For a Lebesgue density f on $\mathbf{R}^d$, with entropy

$$H(f) \,=\, -\int_{\mathbf{R}^d} f(t)\, \log f(t)\, dt\,,$$

we will make use of the following well-known and easily established properties.

Lemma 2.4. **(a)** *For $T \in \mathbf{R}^d$ with density f, $U = \psi(T) \sim g$ with $\psi : \mathbf{R}^d \to \mathbf{R}^d$ invertible with inverse Jacobian J_ψ, we have $H(g) = -\mathbf{E}\log|J_\psi| + H(f)$;*

(b) *Let $T = (T_{(1)}, T_{(2)}) \sim f$ be a random vector with independently distributed components $T_{(1)} \sim f_1$ on $\mathbf{R}^{m_1}$ and $T_{(2)} \sim f_2$ on $\mathbf{R}^{m_2}$. Then (as in Lemma 2.3), we have $H(f) = H(f_1) + H(f_2)$.*

As implied by part **(a)** of the above lemma, the entropy $H(f_\mu)$ is constant as a function of μ for location family densities $f_\mu(t) = f_0(t-\mu)$, as is the case for $MMN_d(\mu, b, \Sigma, \mathcal{L})$ densities. Now, we have the following dimension reduction decomposition for the entropy $H_d(b, \Sigma, \mathcal{L})$ of a $MMN_d(0, b, \Sigma, \mathcal{L})$ density.

Lemma 2.5. *We have for $d \geq 2$:*

$$H_d(b, \Sigma, \mathcal{L}) \,=\, H_1(\sqrt{b^\top \Sigma^{-1} b}, 1, \mathcal{L}) \,+\, \frac{d-1}{2}\,\{1 + \log(2\pi)\} \,+\, \frac{1}{2}\log|\Sigma|\,.$$

Proof. Let $X \sim MMN_d(0, b, \Sigma, \mathcal{L})$, which has entropy $H_d(b, \Sigma, \mathcal{L})$, and set $Z = H\,\Sigma^{-1/2}\,X \sim f_Z$ with H orthogonal having first row $\frac{b^\top \Sigma^{-1/2}}{\sqrt{(b^\top \Sigma^{-1} b)}}$. It follows from part **(a)** of Lemma 2.4 that $H(f_Z) = -\frac{1}{2}\,\log|\Sigma| + H_d(b, \Sigma, \mathcal{L})$. From Lemma 2.1, we have $Z = (Z_1, Z_{(2)})^\top$ with $Z_1 \sim MMN_1(0, \sqrt{(b^\top \Sigma^{-1} b)}, 1, \mathcal{L})$ and $Z_{(2)} \sim N_{d-1}(0, I_{d-1})$ independently distributed, and the result follows from part **(b)** of Lemma 2.4 and a straightforward evaluation of the entropy $H(\phi_{d-1})$. ☐

With the above, we conclude with an expression for the constant and minimax risk of the MRE density.

Theorem 2.2. *In the context of model (2.4), the Kullback-Leibler risk of the MRE density $\hat{q}_U$ is given by*

$$R_{KL}(\theta, \hat{q}_U) \,=\, H_1(\sqrt{a^\top \Sigma_S^{-1} a}, 1, \mathcal{L}_3) \,-\, H_1(\sqrt{a^\top \Sigma_Y^{-1} a}, 1, \mathcal{L}_2) + \frac{1}{2}\log\frac{|\Sigma_S|}{|\Sigma_Y|}\,, \tag{2.9}$$

with $\Sigma_S \,=\, \Sigma_X + \Sigma_Y$.

Proof. Observe first setting $D =^d Y - X$ that

$$\begin{aligned} \mathbf{E}_\theta\, \log \hat{q}_U(Y; X) \quad &= \quad \mathbf{E}_\theta\, \log \hat{q}_U(Y - X; 0) \\ &= \quad \mathbf{E}_0\, \log \hat{q}_U(Y - X; 0) \end{aligned}$$

$$\begin{aligned} &= \mathbf{E}_0 \log \hat{q}_U(D;0) \\ &= -H(\hat{q}_U(\cdot;0))\,, \end{aligned}$$

since $D|\theta \sim MMN_d(0, a, \Sigma_S, \mathcal{L}_3)$, i.e., D has density $\hat{q}_U(\cdot;0)$. Therefore, we obtain for $\theta \in \mathbf{R}^d$ that

$$\begin{aligned} R_{KL}(\theta, \hat{q}_U) &= \mathbf{E}_\theta\{\log q_\theta(Y) - \log \hat{q}_U(Y;X)\} \\ &= H(\hat{q}_U(\cdot;0)) - H(q_0) \\ &= H_d(a, \Sigma_S, \mathcal{L}_3) - H_d(a, \Sigma_Y, \mathcal{L}_2)\,, \end{aligned}$$

and the result then follows from Lemma 2.5. □

The particular case with $\Sigma_X = \sigma_X^2 I_d$ and $\Sigma_Y = \sigma_Y^2 I_d$ follows directly from (2.9) and yields

$$R_{KL}(\theta, \hat{q}_U) = H_1(\frac{\|a\|}{\sigma_S}, 1, \mathcal{L}_3) - H_1(\frac{\|a\|}{\sigma_Y}, 1, \mathcal{L}_2) + \frac{d}{2} \log \frac{\sigma_S^2}{\sigma_Y^2}\,. \tag{2.10}$$

3. Bayes posterior analysis and predictive densities

In this section, given a relative paucity of results in the literature, we expand on representations for Bayesian posterior and predictive densities for mean-mixture of normal distributions. For skewed-normal distributions, previous Bayesian posterior and predictive analysis, with applications of interest to empirical Bayes estimators of θ and to portfolio theory, have appeared in [17] and [24], respectively.

Bayesian posterior analysis of MMN models relate to the general form

$$X|K, \theta \sim f_{\theta,K}\,,\ K \sim g\,,\ \text{ and } \theta \sim \pi\,, \tag{3.1}$$

with observable $X \in \mathbf{R}^d$, density g of K free of θ, and π prior density for $\theta \in \mathbf{R}^d$. Such a set-up leads to the following intermediate result, taken from [21].

Lemma 3.1. *For model (3.1), the posterior distribution of $U =^d \theta|x$ admits the representation*

$$U|K' \sim \pi_{k',x} \text{ with } K' \sim g_{\pi,x}\,, \tag{3.2}$$

$\pi_{k',x}$ being the posterior density of θ as if $K = k'$ had been observed, and $g_{\pi,x}(k') \propto g(k')\, m_{\pi,k'}(x)$ with $m_{\pi,k'}$ being the marginal density of X as if $K = k'$ had been observed.

We now apply the above to MMN distributions.

Example 3.1. *Consider $X|\theta \sim MMN_d(\theta, a, \Sigma, \mathcal{L})$ and the prior $\theta \sim N_d(\mu, \Delta)$ with $\Sigma, \Delta > 0$. The above fits into model (3.1) with g taken to be the density of the mixing parameter $K = V \sim \mathcal{L}$, and $f_{\theta,k}$ the $N_d(\theta + ka, \Sigma)$ density. Conditional on $K = k'$, standard Bayesian analysis for the normal model tells us that*

$$\theta|k', x \sim N_d\big((I-P)x + P\mu - k'a, (I-P)\,\Sigma\big)\,, \text{ and } X|k' \sim N_d(\mu + k'a, \Sigma + \Delta)\,, \tag{3.3}$$

with $P = \Sigma\,(\Sigma+\Delta)^{-1}$, *which yields the densities* $\pi_{k',x}$ *and* $m_{\pi,k'}$ *of Lemma 3.1. Then from Lemma 3.1, we infer that*

$$\theta|x \sim MMN_d\left((I-P)x + P\mu,\; a^* = -a\,,\; (I-P)\,\Sigma\,, \mathcal{L}^*\right)\,, \tag{3.4}$$

where the distribution $\mathcal{L}^*$ *has density*

$$g_{\pi,x}(k') \,\propto g(k')\, e^{-\frac{A}{2}k'^2 + Bk'}\,, \tag{3.5}$$

with $A = a^\top(\Sigma+\Delta)^{-1}a$ *and* $B = (x-\mu)^\top(\Sigma+\Delta)^{-1}a\,.$

Furthermore, it follows immediately that

$$\mathbf{E}(\theta|x) \;=\; (I-P)x + P\mu \;-\; P\,a\,\mathbf{E}(K')\,, \;\; \text{with } K' \sim g_{\pi,x}\,. \tag{3.6}$$

We conclude here by pointing out that a similar analysis was provided by [17] for $\Delta = \tau^2 I_d$ *and a multivariate skew normal model. This was carried out in the context of an empirical Bayesian analysis.*

Remark 3.1. *For the improper prior density* $\pi(\theta) = 1$, *one obtains* $\theta|x \sim MMN_d(x, -a, \Sigma, \mathcal{L})$ *by a direct calculation. It can also be inferred from the above Example with* $\Delta = \tau^2 I_d$ *and* $\tau^2 \to \infty$.

Example 3.2. *It is interesting to further study the above posterior distributions for the particular cases where the mixing density (i.e.,* V *or* K*) of the MMN model is of the form*

$$g(k) \propto e^{-c_1 k^2/2 \,-\, c_2 k}\;\mathbf{I}_{(0,\infty)}(k), \tag{3.7}$$

with $c_1 > 0, c_2 \in \mathbf{R}$, *or* $c_1 = 0, c_2 > 0$. *Several of these distributions are presented in Example 6.1. Cases* $c_1 > 0$ *for instance, which correspond to truncated normal distributions on* $(0,\infty)$, *lead to skew-normal densities (2.3) for* $c_2 = 0$. *In the following, denote* $TN(a, b; (0,\infty))$ *as a truncated normal distribution on* $(0,\infty)$ *with shape parameter* $a \in \mathbf{R}$, *scale parameter* $b > 0$, *density* $\frac{1}{b}\,\frac{\phi((y-a)/b)}{\Phi(a/b)}\,\mathbf{I}_{(0,\infty)}(y)$, *and expectation* $a + bR(a/b)$, *with the reverse Mill's ratio* $R(\cdot)$, *given by* $R(t) = \frac{\phi(t)}{\Phi(t)}, t \in \mathbf{R}$.

Now, for $K \sim g$ *as in (3.7), it follows from (3.5) that*

$$\begin{aligned} g_{\pi,x}(k') \quad &\propto \quad e^{-(c_1+A)k'^2/2 \,+\, (B-c_2)k'}\,\mathbf{I}_{(0,\infty)}(k') \\ &\propto \quad \phi\left(\sqrt{A+c_1}\,k' \;-\; \frac{(B-c_2)}{\sqrt{A+c_1}}\right)\mathbf{I}_{(0,\infty)}(k')\,, \end{aligned}$$

which is the density of a $TN\left(\frac{B-c_2}{A+c_1}, \frac{1}{\sqrt{A+c_1}}; (0,\infty)\right)$ *distribution. Hence, the above, which yields the density associated with* $\mathcal{L}$, *provides a complete description of the posterior distribution in (3.4) for all considered cases of mixing density (3.7). Analogously, the corresponding expectation* $\mathbf{E}(K') = \frac{B-c_2}{A+c_1} + \frac{1}{\sqrt{A+c_1}}\,R(\frac{B-c_2}{\sqrt{A+c_1}})$ *provides an explicit expression for the posterior expectation* $\mathbf{E}(\theta|x)$ *in (3.6).*

3.1. Predictive densities

We now continue the above posterior analysis by focussing on the Bayes predictive density (i.e., the conditional density of Y given $X = x$) for MMN distributions and a normally distributed prior for the unknown location parameter. In doing so, the following extension comes into play.

Definition 3.1. *A random vector $Z \in \mathbf{R}^d$ is said to have a mean-mixture of normal distribution with two directions, denoted as $Z \sim MMN_d(\theta, a_1, a_2, \Sigma, \mathcal{L})$, if it admits the representation*

$$Z|W_1, W_2 \sim N_d\left(\theta + a_1 W_1 + a_2 W_2\,, \Sigma\right) \text{ with } (W_1, W_2) \sim \mathcal{L},$$

where $\theta \in \mathbf{R}^d$ is a location parameter, $a_1, a_2 \in \mathbf{R}^d$ are known perturbation vectors, Σ is a known positive definite matrix, and W_1, W_2 are scalar random variables with joint cdf $\mathcal{L}$.

We make use of the following intermediate result provided in [21] and applicable to mixture models of the form:

$$X|K, \theta \sim f_{\theta,K} \text{ with } K \sim g\,; Y|J, \theta \sim f_{\theta,J} \text{ with } J \sim h, \text{ and } \theta \sim \pi. \tag{3.8}$$

In the above set-up, $X \in \mathbf{R}^d$ is observable, the mixing variables K and J are independently distributed with distributions free of θ, the variables X and Y are conditionally independent on θ, and π is a prior density for $\theta \in \mathbf{R}^d$ with respect to a σ-finite measure ν.

Lemma 3.2. *For model (3.8), setting $\pi_{k',x}$ and $g_{\pi,x}$ as in Lemma 3.1, the Bayes predictive density of Y admits the mixture representation*

$$Y|J', K' \sim q_\pi(\cdot|J', K'), \text{ with } J' \sim h, K' \sim g_{\pi,x} \text{ independent },$$

and $q_\pi(y|j', k') = \int_{\mathbf{R}^d} q_{\theta,j'}(y)\, \pi_{k',x}(\theta)\, d\nu(\theta)$, which can be interpreted as the Bayes predictive density for Y as if $Y \sim q_{\theta,j'}$ and $K = k'$ had been observed.

When applying the above to mean-mixture of multivariate normal distributions with a normal distributed prior, we obtain the following.

Theorem 3.1. **(a)** *For $X|\theta \sim MMN_d(\theta, a_X, \sigma_X^2 I_d, \mathcal{L}_1)$, $Y|\theta \sim MMN_d(\theta, a_Y, \sigma_Y^2 I_d, \mathcal{L}_2)$ independent, and prior $\theta \sim N_d(\mu, \tau^2 I_d)$, the Bayes predictive distribution for Y is*

$$MMN_d\left(\omega x + (1-\omega)\mu, -\omega a_X, a_Y, (\omega\sigma_X^2 + \sigma_Y^2) I_d, \mathcal{L}\right),$$

with $\mathcal{L}$ the joint cdf of (K', J') for independently distributed $K' \sim g_{\pi,x}$ as in (3.5) and $J' \sim \mathcal{L}_2$, with $\omega = \tau^2/(\tau^2 + \sigma_X^2)$, $A = \|a_X\|^2/(\sigma_X^2 + \tau^2)$, and $B = \{(x-\mu)^\top a_X\}/\{\sigma_X^2 + \tau^2\}$.

(b) *Moreover, whenever $a_Y = c a_X$ for $a_X \neq 0$ and a fixed $c \in \mathbf{R}$, the above predictive distribution is $MMN_d\left(\omega x + (1-\omega)\mu, a_X, (\omega\sigma_X^2 + \sigma_Y^2) I_d, \mathcal{L}_3\right)$, with $\mathcal{L}_3$ the cdf of $cJ' - \omega K'$, and (J', K') distributed as above. Finally, for $a_X = 0$, i.e., for $X|\theta \sim N_d(\theta, \sigma_X^2 I_d)$, the predictive distribution is $MMN_d(\omega x + (1-\omega)\mu, a_Y, (\omega\sigma_X^2 + \sigma_Y^2) I_d, \mathcal{L}_2)$.*

Proof. Part **(b)** follows immediately from part **(a)**. For part **(a)**, consider $X' = X - K'a_X$ and $Y' = Y - J'a_Y$. The result then follows from Lemma 3.2 with the familiar predictive density estimation result:

$$Y'|J', K', X' \sim N_d\big(\omega X' + (1-\omega)\mu, (\omega\sigma_X^2 + \sigma_Y^2)I_d\big),$$

implying

$$q_\pi(\cdot|J', K') \sim N_d\left(\omega x + (1-\omega)\mu - \omega a_X K' + a_Y J', (\omega\sigma_X^2 + \sigma_Y^2)I_d\right),$$

matching Definition 3.1 with $(W_1, W_2) =^d (K', J')$. □

Observe that in the context above where $X|\theta$ is normally distributed, the model and predictive densities for Y are interestingly in the same MMN class with differences only in location and scale, but not with regards to the perturbation factor and mixing distribution.

Remark 3.2. *We point out that the minimum risk predictive density matches the density in* **(b)** *with $\tau^2 = \infty$, i.e., $\omega = 1$.*

4. Dominance results

In this section, we first provide KL risk improvements on the MRE predictive density $\hat{q}_U$ for estimating the density of $Y|\theta \sim MMN_d(\theta, a, \Sigma_Y, \mathcal{L}_2)$ based on $X|\theta \sim MMN_d(\theta, a, \Sigma_X, \mathcal{L}_1)$ with $d \geq 4$. Such improvements are necessarily minimax as a consequence of Lemma 2.1. Our findings cover two types of improvements: **(i)** plug-in type (Section 4.1), and **(ii)** Bayesian improvements (Section 4.2). Furthermore, we provide in Section 4.3 analogous results for certain type of restricted parameter spaces which are also applicable for $d = 2, 3$. Examples will be provided in Section 5.

4.1. Plug-in type improvements

In the normal case with $X|\theta \sim N_d\left(\theta, \sigma_X^2 I_d\right)$ and $Y|\theta \sim N_d\left(\theta, \sigma_Y^2 I_d\right)$ independently distributed, the MRE predictive density $\hat{q}_U(\cdot; X) \sim N_d\left(X, (\sigma_X^2 + \sigma_Y^2)I_d\right)$ is inadmissible for $d \geq 3$ and can be improved for instance by plug-in type densities of the form $q_{\hat{\theta}}(\cdot; X) \sim N_d\left(\hat{\theta}(X), (\sigma_X^2 + \sigma_Y^2)I_d\right)$. Indeed, the KL risk performance of $q_{\hat{\theta}}$ relates directly to the "dual" point estimation risk of $\hat{\theta}(X)$ for estimating θ under squared error loss $\|\hat{\theta} - \theta\|^2$, with $q_{\hat{\theta}}(\cdot; X)$ dominating $\hat{q}_U(\cdot; X)$ if and only if $\hat{\theta}(X)$ dominates X ([10]). For MMN distributions, such a duality does not deploy itself in the same way, but does so after a transformation to a canonical form. In the following, we consider plug-in type densities of the form $q_{\hat{\theta}}(\cdot; X) \sim MMN_d\left(\hat{\theta}(X), a, \Sigma_X + \Sigma_Y, \mathcal{L}_3\right)$, which include $\hat{q}_U$ for $\hat{\theta}(X) = X$, and provide sufficient conditions for which $q_{\hat{\theta}}$ dominates $\hat{q}_U$. The result applies for $d \geq 4$ and for $\Sigma_Y = c\Sigma_X$, but is otherwise general.

Theorem 4.1. *Let X, Y be distributed as in model (2.4) with $a \neq 0, d \geq 4, \theta \in \mathbf{R}^d, \Sigma_Y = c\Sigma_X,$ and set $\Sigma_S = \Sigma_X + \Sigma_Y$. Consider the problem of obtaining a predictive density estimator $\hat{q}(y;X)$, $y \in \mathbf{R}^d$, for the density of Y. Then, $q_{\hat{\theta}}$ dominates $\hat{q}_U$ under KL loss whenever $\hat{\theta}(X) = \Sigma_S^{1/2} H^\top \begin{pmatrix} h_1^\top \Sigma_S^{-1/2} X \\ \hat{\eta}(H_2 \Sigma_S^{-1/2} X) \end{pmatrix}$ with*

- *(i) $H = \begin{pmatrix} h_1^\top \\ H_2 \end{pmatrix}$ a $d \times d$ orthogonal matrix such that $h_1^\top = \frac{a^\top \Sigma_S^{-1/2}}{\sqrt{a^\top \Sigma_S^{-1} a}}$, so that $H\Sigma_S^{-1/2} a = (\sqrt{a^\top \Sigma_S^{-1} a}, 0, \ldots, 0)^\top = a_0 (say)$,*
- *(ii) $\hat{\eta}(Z)$ an estimator of $\eta \in \mathbf{R}^{d-1}$ which dominates Z under squared error loss $\|\hat{\eta} - \eta\|^2$ and for the model $Z|\eta \sim N_{d-1}\left(\eta, (1+c)^{-1} I_{d-1}\right)$.*

Before outlining a proof of the theorem, we highlight its essential nature, which is the provide a recipe for obtaining when $d \geq 4$ an improved KL risk predictive density over the MRE density $\hat{q}_U$. The key idea is to decompose the prediction problem into two separate parts which are additive in terms of risk: a one-dimensional part whose risk contribution matches that of that of $\hat{q}_U$ and another $d-1$ dimensional part, independent of both the perturbation vector and the mixing distributions, for which improvement is possible by relying on well-studied shrinkage type estimators in normal models. Given the generality of the result, a substantial number of applications follow.

Proof of Theorem 4.1. The KL risk difference between $\hat{q}_U$ and $q_{\hat{\theta}}$ is

$$
\begin{aligned}
\Delta_{KL}(\theta, q_{\hat{\theta}}) &= R_{KL}(\theta, q_{\hat{\theta}}) - R_{KL}(\theta, \hat{q}_U) \\
&= \mathbf{E} \log \frac{\int \phi_d\big(\Sigma_S^{-1/2}(Y - X - va)\big)\, d\mathcal{L}_3(v)}{\int \phi_d\big(\Sigma_S^{-1/2}(Y - \hat{\theta}(X) - va)\big)\, d\mathcal{L}_3(v)} \\
&= \mathbf{E} \log \frac{\int \phi_d(H\Sigma_S^{-1/2}(Y - X - va))\, d\mathcal{L}_3(v)}{\int \phi_d(H\Sigma_S^{-1/2}(Y - \hat{\theta}(X) - va))\, d\mathcal{L}_3(v)} \\
&= \Bigg\{ \mathbf{E} \log \frac{\int \phi_1\big(h_1^\top \Sigma_S^{-1/2}(Y - X) - v\sqrt{a^\top \Sigma_S^{-1} a}\big)\, d\mathcal{L}_3(v)}{\int \phi_1\big(h_1^\top \Sigma_S^{-1/2}(Y - \hat{\theta}(X)) - v\sqrt{a^\top \Sigma_S^{-1} a}\big)\, d\mathcal{L}_3(v)} \\
&+ \mathbf{E} \log \frac{\phi_{d-1}\big(H_2 \Sigma_S^{-1/2}(Y - X)\big)}{\phi_{d-1}\big(H_2 \Sigma_S^{-1/2}(Y - \hat{\theta}(X))\big)} \Bigg\}.
\end{aligned}
$$

Now for $\hat{\theta}(X)$ given, the first term above vanishes and we get

$$
\begin{aligned}
\Delta_{KL}(\theta, q_{\hat{\theta}}) &= \mathbf{E} \log \frac{\phi_{d-1}\big(H_2 \Sigma_S^{-1/2}(\theta - X)\big)}{\phi_{d-1}\big(H_2 \Sigma_S^{-1/2}(\theta - \hat{\theta}(X))\big)} \\
&= \frac{1}{2} \left\{ \mathbf{E}\|\hat{\eta}(Z) - \eta\|^2 - \mathbf{E}\|Z - \eta\|^2 \right\},
\end{aligned}
\tag{4.1}
$$

for $Z := H_2 \Sigma_S^{-1/2} X \sim N_{d-1}\left(\eta, (1+c)^{-1} I_{d-1}\right)$ and $\eta := H_2 \Sigma_S^{-1/2} \theta$. □

The above dominance finding is quite general with respect to the specifications of $a, \mathcal{L}_1$, and $\mathcal{L}_2$ of model (2.4). Furthermore, observe by examining (4.1) that the risk difference depends on θ only through $\eta = H_2 \Sigma_S^{-1/2}\theta$ and this for any choice of H_2. More strikingly as seen with (4.1), the risk difference does not depend on the mixing distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ and can be simply described by a quadratic risk difference of point estimators in a $(d-1)$ variate normal distribution problem. An illustration of Theorem 4.1 and some of the above features will be presented in Section 5, but we conclude first with an application of Theorem 4.1 to Baranchik-type estimators.

Corollary 4.1. *Under the set-up of Theorem 4.1, a predictive density* $q_{\hat{\theta}_{k,r}}$ *dominates* $\hat{q}_U$ *under KL loss for*

$$\hat{\theta}_{k,r}(X) = X - k(1+c)\frac{r\left(\frac{\|H_2\Sigma_X^{-1/2}X\|^2}{1+c}\right)}{\|H_2\Sigma_X^{-1/2}X\|^2}\Sigma_X^{1/2}H_2^\top H_2\Sigma_X^{-1/2}X, \tag{4.2}$$

whenever $r(\cdot)$ *is an absolutely continuous and nondecreasing function on* $(0,\infty)$ *such that: (i)* $0<k<2(d-3)$ *and (ii)* $0 \le r(\cdot) \le 1$.

Proof. It follows from a well-known dominance condition (e.g., Theorem 2.3 in [11]) that the Baranchik-type estimator $\left(1-(1+c)^{-1}\frac{r(\|Z\|^2)}{\|Z\|^2}\right)Z$ dominates Z under loss $\|\hat{\eta}-\eta\|^2$ for $Z \sim N_{d-1}\left(\eta,(1+c)^{-1}I_{d-1}\right)$, $d \ge 4$, and under conditions (i) and (ii). The result then follows as an application of Theorem 4.1 upon expressing $\hat{\theta}_{k,r}(X)$ in terms of $\Sigma_X = (1+c)^{-1}\Sigma_S$. □

4.2. Bayesian improvements

We now focus on Bayesian predictive densities that dominate $\hat{q}_U$. In doing so, we work with Lemma 2.1's canonical form, apply the partitioning argument of Lemma 2.3, and take advantage of known results for prediction in multivariate normal models. We consider a class of improper priors on θ which is the product measure of a (improper) uniform density over the linear subspace spanned by a and a second component of the prior (π_0) supported on the subspace orthogonal to a. The measure of this nature splits resulting Bayes predictive densities into independent parts, and leads to a decomposition of the KL risk in two additive parts. The dominance result is thus derived by dominating the part of the KL risk corresponding to the orthogonal space to a, where transformed variables are N_{d-1} distributed, and where we can capitalize on known results. Namely, the superharmonicity of π_0, or its associated marginal density or its associated square root marginal density, will suffice for dominance and minimaxity. The following is stated for $\Sigma_X = \sigma_X^2 I_d$ and $\Sigma_Y = \sigma_Y^2 I_d$, but the finding applies as well to cases where $\Sigma_Y = c\Sigma_X$ (see Remark 6.1 in the Appendix).

We choose an orthogonal transformation $H = \begin{pmatrix} h_1^\top \\ H_2 \end{pmatrix}$ of variables to obtain following results such that $h_1 = \frac{a}{\|a\|}$. We denote $\zeta = H_2\theta \in \mathbf{R}^{d-1}$, where the rows of H_2 span the $(d-1)$-dimensional space orthogonal to a. The following intermediate result demonstrates the splitting of Bayesian predictive densities associated with the class of priors considered here.

Lemma 4.1. *Consider X, Y distributed as in model (2.4) with $\Sigma_X = \sigma_X^2 I_d$, $\Sigma_Y = \sigma_Y^2 I_d$, and $d \geq 2$. Consider prior densities of the form*

$$\pi(\theta) = \pi_0\left(\zeta\right) = \pi_0\left(H_2\theta\right). \tag{4.3}$$

Then, the Bayes predictive density for Y is given by

$$\hat{q}_\pi(y; X) \;=\; \hat{q}_U(h_1^\top y; h_1^\top X) \times \hat{q}_{\pi_0}(H_2 y; H_2 X)\,, \tag{4.4}$$

with: **(i)** *$\hat{q}_U(\cdot; h_1^\top X)$ the MRE density, given in Theorem 2.1, of $h_1^\top y \sim MMN_1(h_1^\top\theta, \|a\|, \sigma_Y^2, \mathcal{L}_2)$ based on $h_1^\top X \sim MMN_1(h_1^\top\theta, \|a\|, \sigma_X^2, \mathcal{L}_1)$, and* **(ii)** *$\hat{q}_{\pi_0}(\cdot; H_2X)$ the Bayes predictive density for $H_2Y \sim N_{d-1}(\zeta, \sigma_Y^2 I_{d-1})$ based on $H_2X \sim N_{d-1}\left(\zeta, \sigma_X^2 I_{d-1}\right)$ and for prior density $\pi_0(\zeta)$ for ζ.*

Proof. From the transformation of variables under the orthogonal matrix H, it follows that $\hat{q}_\pi(y; X) \;=\; \hat{q}_\pi(Hy; HX)$. Note that the distribution of the transformed variables is

$$\begin{aligned} &HX \sim MMN_d(H\theta, a_0, \sigma_X^2 I_d, \mathcal{L}_1) \\ \text{and} \quad &HY \sim MMN_d(H\theta, a_0, \sigma_Y^2 I_d, \mathcal{L}_2), \end{aligned}$$

where $a_0 = (\|a\|, 0, \ldots, 0)^\top$. With the multiplicative nature of the prior density in (4.3) in terms of $h_1^\top\theta$ and $\zeta = H_2\theta$, coupled with the conditional independence of $h_1^\top Y$ and H_2Y given θ, expression (4.4) follows. ☐

Now we provide the main dominance result of this section by establishing parallels between predictive density estimation problems for d-variate MMN model and $(d-1)$- variate normal model. We denote, for a given prior density π_0, m_{π_0} as the marginal density of $Z \sim N_{d-1}(\zeta, \sigma^2 I_{d-1})$. It is given by

$$m_{\pi_0}(z, \sigma^2) = \int_{\mathbf{R}^{d-1}} \phi_{d-1}(z-\zeta, \sigma^2 I_{d-1})\, \pi_0(\zeta)\, d\zeta\,. \tag{4.5}$$

Theorem 4.2. *Consider X, Y distributed as in model (2.4) with $\Sigma_X = \sigma_X^2 I_d$, $\Sigma_Y = \sigma_Y^2 I_d$, and $d \geq 2$, and consider prior densities $\pi(\theta)$ of the form given in (4.3).*

- **(a)** *Then, the corresponding Bayes predictive density $\hat{q}_\pi$ admits the representation*
$$\hat{q}_\pi(y; x) = \hat{q}_U(y; x) \times \frac{m_{\pi_0}(H_2\, w, \sigma_W^2)}{m_{\pi_0}(H_2\, x, \sigma_X^2)}\,, \tag{4.6}$$
with $w = \frac{\sigma_X^2 y + \sigma_Y^2 x}{\sigma_X^2 + \sigma_Y^2}$, and $\sigma_W^2 = \frac{\sigma_X^2 \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}$.
- **(b)** *If $d \geq 4$, then $\hat{q}_\pi$ given in (4.6) dominates the MRE density $\hat{q}_U$ under Kullback-Leibler loss, and is therefore minimax, if and only if $\hat{q}_{\pi_0}(\cdot; H_2X)$ dominates under Kullback-Leibler loss the MRE density for H_2Y based on H_2X and given by a $N_{d-1}\big(H_2X, (\sigma_X^2 + \sigma_Y^2)I_{d-1}\big)$ density.*

Proof. **(a)** For the multivariate normal case with $H_2X \sim N_{d-1}(\zeta, \sigma_X^2 I_{d-1})$ independent of $H_2Y \sim N_{d-1}(\zeta, \sigma_Y^2 I_{d-1})$, it is known (i.e., Lemma 2 in [13]) that the Bayes predictive density associated with prior density π_0 for ζ is given by

$$\hat{q}_{\pi_0}(H_2y; H_2X) = \hat{q}_U(H_2y; H_2X) \times \frac{m_{\pi_0}(H_2w; \sigma_W^2)}{m_{\pi_0}(H_2x; \sigma_X^2)}, \tag{4.7}$$

where $\hat{q}_U(\cdot; H_2X)$ is the MRE predictive density of the density of H_2Y based on H_2X, and given by a $N_{d-1}\big(H_2X, (\sigma_X^2 + \sigma_Y^2)\, I_{d-1}\big)$ density. Now, plugging the above (4.7) into (4.4), we obtain

$$\hat{q}_\pi(y; X) \;=\; \hat{q}_U(h_1^\top y; h_1^\top X) \times \hat{q}_U(H_2y; H_2X) \times \frac{m_{\pi_0}(H_2w; \sigma_W^2)}{m_{\pi_0}(H_2x; \sigma_X^2)},$$

and the result follows since $m_{\pi_0}(z; \sigma^2) \;=\; 1$ for the uniform density $\pi_0 = 1$.

(b) It immediately follows from **(a)** and (4.7) that

$$\begin{aligned} R_{KL}(\theta, \hat{q}_U) - R_{KL}(\theta, \hat{q}_\pi) &= \mathbf{E} \log m_{\pi_0}(H_2W; \sigma_W^2) - \mathbf{E} \log m_{\pi_0}(H_2X; \sigma_X^2) \\ &= \mathbf{E} \log \hat{q}_{\pi_0}(H_2Y; H_2X) - \mathbf{E} \log \hat{q}_U(H_2Y; H_2X). \end{aligned}$$

Observe that the r.h.s. is indeed the KL risk difference between the MRE density for H_2Y based on H_2X and $\hat{q}_{\pi_0}(\cdot; H_2X)$, the Bayes predictive density of H_2Y under the prior density $\pi_0(\zeta)$. This yields part (**b**). □

Remark 4.1. *Theorem 4.2's dominance finding in part* **(b)** *is unified with respect to the model settings a, $\mathcal{L}_1$ and $\mathcal{L}_2$, as well as the dimension $d \geq 4, \sigma_X^2$, and σ_Y^2. Furthermore, as seen in the lines of the proof, the difference in risks between the predictive densities $\hat{q}_U$ and $\hat{q}_\pi$: (i) does not depend on the mixing $\mathcal{L}_1$ and $\mathcal{L}_2$, and (ii) depends on θ only through $\zeta = H_2\theta$.*

Starting with [14], continuing namely with [13], several Bayesian predictive densities $\hat{q}_{\pi_0}(\cdot; H_2X)$ have been shown to satisfy the dominance condition in part **(b)** of the above Theorem. Such choices lead to dominating predictive densities of $\hat{q}_U$. In [13], analogously to the quadratic risk estimation problem with multivariate normal observables (e.g., [27, 11]), sufficient conditions for minimaxity are conveniently expressed in terms of the marginal density $m_{\pi_0}(z, \sigma^2)$ given in (4.5). The superharmonicity of either π_0, $m_{\pi_0}(z, \sigma^2)$ for $z \in \mathbf{R}^{d-1}$, for various values of σ^2, or as well of $\sqrt{m_{\pi_0}(z, \sigma^2)}$, each lead to sufficient conditions for minimaxity. We recall here that the superharmonicity of $h : \mathbf{R}^{d-1} \to \mathbf{R}$ holds whenever the Laplacian $\Delta^2 h(t) = \sum_{i=1}^{d-1} \frac{\partial^2 h(t)}{\partial t_i^2}$ exists with $\Delta^2 h(t) \leq 0$ for $t \in \mathbf{R}^{d-1}$.

Corollary 4.2. *Consider the prediction context of Theorem 4.2 and a prior density π_0 as in (4.3) other than the uniform density. Suppose that $m_{\pi_0}(z, \sigma_X^2)$ is finite for all $z \in \mathbf{R}^{d-1}$ and that $d \geq 4$. Then, the following conditions are each sufficient for $\hat{q}_\pi(\cdot; X)$ given in (4.6) to dominate the MRE density $\hat{q}_U$:*

(i) $\Delta^2\, m_{\pi_0}(z,\sigma^2) \,\le\, 0$, $z \in \mathbf{R}^{d-1}$, *for* $\frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2} < \sigma^2 < \sigma_X^2$, *with strict inequality on a set of positive Lebesgue measure on* $\mathbf{R}^{d-1}$ *for at least one* σ^2;

(ii) $\Delta^2 \sqrt{m_{\pi_0}(z,\sigma^2)} \,\le\, 0$, $z \in \mathbf{R}^{d-1}$, *for* $\frac{\sigma_X^2\sigma_y^2}{\sigma_X^2+\sigma_Y^2} < \sigma^2 < \sigma_X^2$, *with strict inequality on a set of positive Lebesgue measure on* $\mathbf{R}^{d-1}$ *for at least one* σ^2;

(iii) *The prior* π_0 *is such that* $\Delta^2 \pi_0(\zeta) \le 0$ *a.e.*

Proof. The results follow from part (b) of Theorem 4.2 and Theorem 1 – Corollary 2 in [13]. □

Choices of the prior density π_0 satisfying the conditions of Corollary 4.2 thus rest upon analyses for the normal case which are plentiful. In particular, several examples of π_0, and the resulting predictive density $\hat{q}_{\pi_0}$, are provided in [13]. These provide explicit representations of minimax predictive densities $\hat{q}_\pi$ given in (4.6). A detailed example is presented in Section 5.

To conclude describing the dominance findings of this section and of Section 4.1, we point out that the plug-in type improvements of Theorem 4.2 and the Bayesian dominance results of Theorem 4.2 and Corollary 4.2 are applicable regardless of the choice of the orthogonal completion H_2 of H, thus adding to choices of π_0 leading to minimaxity. Furthermore, the above developments are unified and the findings are applicable for all MMN models (2.4) with $\Sigma_X = \sigma_X^2 I_d$ and $\Sigma_Y = \sigma_Y^2 I_d$, as well as for $\Sigma_Y = c\Sigma_X$ as justified in Remark 6.1.

Remark 4.2. *A particular appealing choice of* H_2*, which will be further explored below in Sections 4.3 and 5, is such that* $H_2^\top H_2 = I_d - \frac{aa^\top}{a^\top a}$ *in which case*

$$\|\zeta\|^2 \,=\, \theta^\top \left(I_d - \frac{aa^\top}{a^\top a} \right) \theta\,, \tag{4.8}$$

and spherically symmetric densities $\pi_2(\zeta) \,=\, g\left(\|\zeta\|^2\right)$ *leading to prior densities in (4.3) of the form*

$$\pi(\theta) \,=\, g\left\{ \theta^\top \left(I_d - \frac{aa^\top}{a^\top a} \right) \theta \right\} \,=\, g\left(\left\| \theta - \frac{a^\top \theta}{a^\top a}\, a \right\|^2 \right). \tag{4.9}$$

Such densities do not depend on $\|a\|$ and have contours given by hypersurfaces of cylinders with axis given by a (or $h_1 = \frac{a}{\|a\|}$). Here is an example of three contours for $d = 3$ and $a = (1,1,1)^\top$.

4.3. Restricted parameter spaces

Theorem 4.2 also leads to implications when there exists parametric restrictions on $\zeta = H_2\theta$. Statistical models where parametric restrictions are present appear naturally in a great variety of contexts, and there is a large literature on related inferential problems, namely for a decision-theoretic approach (e.g., [23, 28]).

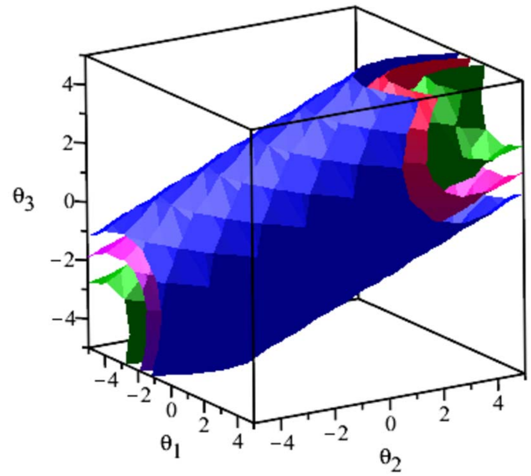


FIG 1. *Contours of $\pi(\theta)$ for $d = 3$ and $a = (1, 1, 1)^\top$.*

Questions of predictive analysis under parametric restrictions are also of interest with findings obtained in [22, 16, 10]. Namely, for normal models, specifically model (2.4) with $a = 0$, $\Sigma_X = \sigma_X^2 I_d$, $\Sigma_Y = \sigma_Y^2 I_d$ with θ constrained to a convex set C with non-empty interior, it was shown in [10] that the Bayes predictive density associated with the uniform prior for θ on C dominates the MRE predictive density under Kullback-Leibler loss. The next result extends this finding to MMN models.

Theorem 4.3. *Consider X, Y distributed as in model (2.4) with $\Sigma_X = \sigma_X^2 I_d$, $\Sigma_Y = \sigma_Y^2 I_d$, and $d \geq 2$. Let $C \subset \mathbf{R}^{d-1}$ be a convex set with non-empty interior, and let $\pi_C(\theta) = \pi_{0,U}(\zeta) = I_C(\zeta)$, with I_C the indicator function of the set C. Then $\hat{q}_{\pi_C}(\cdot; X)$ dominates $\hat{q}_U(\cdot; X)$ under KL risk and the restriction $\theta \in \{\theta \in \mathbf{R}^d : H_2\theta \in C\}$.*

Proof. As in Theorem 4.2 and the given proof, we infer that $\hat{q}_\pi$ given in (4.6) with prior density $\pi(\theta) = \pi_0(\zeta)$ for $\zeta = H_2\theta$ dominates $\hat{q}_U$ if and only if $\hat{q}_{\pi_0}(\cdot; H_2X)$ dominates the MRE density for $H_2Y \sim N_{d-1}(\zeta, \sigma_Y^2 I_{d-1})$. But, since this latter dominance holds precisely for density $\pi = \pi_C$ for the uniform density choice $\pi_0 = \pi_{0,U}$ as shown in [10], the result follows. $\square$

The setting of C above is quite general and interesting examples includes balls, and cones such as order constraints of the type $\zeta_i \leq \zeta_{i+1}$ for $i = 1, \ldots, d-2$. As earlier, the finding is unified and general to the MMN models. Here are two applications of Theorem 4.3.

Example 4.1. *Suppose $d = 2$, $a = (1, 1)^\top$, and the parametric restriction $\underline{c} \leq \theta_1 - \theta_2 \leq \bar{c}$, with $C = (\underline{c}, \bar{c})$ a strict subset of $\mathbf{R}$. The MRE density $\hat{q}_U(\cdot; X)$ is that of $MMN_2\big(X, a, (\sigma_X^2 + \sigma_Y^2) I_2, \mathcal{L}_3\big)$ distribution. In the context of Theorem 4.3, we have $\zeta = \frac{\theta_1 - \theta_2}{\sqrt{2}}$ and the prior density $\pi_C(\theta) = I_C(\theta_1 - \theta_2)$. Theorem 4.2 tells* *us that the Bayes predictive density $\hat{q}_{\pi_C}$ dominates the MRE $\hat{q}_U$ with respect to KL loss and under the given parametric restriction.* [1]

An explicit expression for $\hat{q}_{\pi_C}$ is available from Theorem 4.2 with π_0 the uniform $U(\frac{\underline{c}}{\sqrt{2}}, \frac{\bar{c}}{\sqrt{2}})$ density for ζ. As evaluated in [16], we obtain

$$\left(\frac{\sqrt{2}}{\bar{c}-\underline{c}}\right) m_{\pi_0}(z,\sigma^2) = \int_{\underline{c}/\sqrt{2}}^{\bar{c}/\sqrt{2}} \phi\left(z-\zeta,\sigma^2\right)\, d\zeta = \Phi\left(\frac{z+\bar{c}/\sqrt{2}}{\sigma}\right) - \Phi\left(\frac{z+\underline{c}/\sqrt{2}}{\sigma}\right),$$

and (4.6) then yields

$$\hat{q}_{\pi_C}(y;x) = \hat{q}_U(y;x)\ \frac{\Phi\left(\frac{w+\bar{c}/\sqrt{2}}{\sigma_W}\right) - \Phi\left(\frac{w+\underline{c}/\sqrt{2}}{\sigma_W}\right)}{\Phi\left(\frac{x+\bar{c}/\sqrt{2}}{\sigma_X}\right) - \Phi\left(\frac{x+\underline{c}/\sqrt{2}}{\sigma_X}\right)}\,, y \in \mathbf{R},$$

with w and σ_W^2 given in (4.6), and $\hat{q}_U$ the MRE density which is that of a $MMN_2(x, a, (\sigma_X^2+\sigma_Y^2)I_2, \mathcal{L}_3)$ distribution.

Example 4.2. *Theorem 4.3 applies for θ restricted to a cylinder of radius, say m, with the axis along the direction a, i.e.,*

$$C_m = \left\{\theta \in \mathbf{R}^d : \left\|\theta - \frac{a^\top\theta}{a^\top a}a\right\| \le m\right\};$$

examples of which are drawn in Figure 1. The dominating predictive density $\hat{q}_{\pi_{C_m}}$ is Bayes with respect to the uniform prior density on C_m, which corresponds to (4.9) with $g(t) = I_{(0,m)}(t)$. An explicit expression for $\hat{q}_{\pi_{C_m}}$ can be derived from Theorem 4.2 with π_0 the uniform density on the ball $B_m = \{t \in \mathbf{R}^{d-1} : \|t\| \le m\}$ and marginal density

$$m_{\pi_0}(z,\sigma^2) = \int_{B_m} \phi_{d-1}\left(z-\zeta, \sigma^2 I_{d-1}\right)\, d\zeta = F_{d-1,\frac{\|z\|^2}{\sigma^2}}\left(\frac{m^2}{\sigma^2}\right),$$

with $F_{\nu,\lambda}$ the cdf of a $\chi^2_\nu(\lambda)$ distribution. From (4.6), we thus obtain

$$\hat{q}_{\pi_{C_m}}(y;x) = \hat{q}_U(y;x)\left(\frac{F_{d-1,\frac{\|H_2 w\|^2}{\sigma_W^2}}\left(\frac{m^2}{\sigma_W^2}\right)}{F_{d-1,\frac{\|H_2 x\|^2}{\sigma_X^2}}\left(\frac{m^2}{\sigma_X^2}\right)}\right), y \in \mathbf{R}^d,$$

with $\|H_2 t\|^2 = t^\top\left(I - \frac{aa^\top}{a^\top a}\right)t$, for $t \in \mathbf{R}^d$, $w = \frac{\sigma_X^2 y + \sigma_Y^2 x}{\sigma_X^2+\sigma_Y^2}$, $\sigma_W^2 = \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}$, and $\hat{q}_U$ the MRE density which is that of a $MMN_d(x, a, (\sigma_X^2+\sigma_Y^2)I_d, \mathcal{L}_3)$ distribution.

[1] In Example 4.1, for the compact interval case say without loss of generality $\underline{c} = -m$ and $\bar{c} = m$, there exists a much larger class of dominating predictive densities obtained by replacing the uniform density for ζ by an even density π_0 supported on $(-m, m)$ that is increasing and logconcave on $(0, m)$. This is established as in Theorem 4.3 and making use of Theorem 3.2 in [10], which exploits a related point estimation finding in [18].

5. Illustrations

We provide here illustrations of Theorems 4.1 and 4.2 accompanied by numerical comparisons and various observations.

Example 5.1 (A Bayesian minimax predictive density)**.** *In the context of Theorem 4.2, consider H_2 as in Remark 4.2 combined with the harmonic prior density for $\zeta \in \mathbf{R}^{d-1}$ given by $\pi_0(\zeta) = \|\zeta\|^{-(d-3)}$ and which generates via (4.9) an "adjusted" harmonic prior density on θ given by*

$$\pi_H(\theta) = \left\| \theta - \frac{a^\top \theta}{a^\top a} a \right\|^{-(d-3)} . \tag{5.1}$$

Thus, the prior density is the product measure on $\mathbf{R}^d$ with uniform prior on the linear subspace spanned by a and the above harmonic measure on the $(d-1)$-dimensional chosen subspace orthogonal to a. Since π_0 is superharmonic on $\mathbf{R}^{d-1}$ for $d \geq 4$, it follows from Corollary 4.2 that the Bayes predictive density $\hat{q}_{\pi_H}(\cdot; X)$ in (4.6), as well as in (5.3) below, dominates the MRE density $\hat{q}_U$ and is consequently minimax.

An explicit expression for $\hat{q}_{\pi_H}$ is available from(4.6) with marginal density

$$m_{\pi_0}(z, \sigma^2) = \int_{\mathbf{R}^{d-1}} \phi_{d-1}(z-\zeta, \sigma^2 I_{d-1}) \frac{1}{||\zeta||^{(d-3)}} \, d\zeta = \sigma^{3-d} \, \mathbf{E} \, T^{\frac{(3-d)}{2}},$$

where $T \sim \chi^2_{d-1}\left(\frac{||z||^2}{\sigma^2}\right)$. In particular for odd $d \geq 5$, as shown in the Appendix, one may obtain

$$m_{\pi_0}(z, \sigma^2) = \left(||z||^2\right)^{\frac{3-d}{2}} \left(1 - e^{-\frac{||z||^2}{2\sigma^2}} \sum_{k=0}^{\frac{d-5}{2}} \left(\frac{||z||^2}{2\sigma^2}\right)^k \frac{1}{k!}\right) = s(||z||^2, \sigma^2) \;\; \textit{(say)} \, , \tag{5.2}$$

which relates to expressions for the inverse moments of a chi-square variable with even degrees of freedom (e.g., [8]), as well a closed form for an incomplete gamma function which intervenes in Komaki's [14] representation of m_{π_0}. From (4.6) and the above, we thus have

$$\hat{q}_{\pi_H}(y; x) \; = \hat{q}_U(y; x) \, \frac{s\left(\left\| w - \frac{a^\top w}{a^\top a} a \right\|^2, \sigma_W^2\right)}{s\left(\left\| x - \frac{a^\top x}{a^\top a} a \right\|^2, \sigma_X^2\right)} \, , y \in \mathbf{R}^d \, , \tag{5.3}$$

where w and σ_W^2 are as given in (4.6).

Risk differences between $\hat{q}_U$ and $\hat{q}_{\pi_H}$ are plotted in **Figure 2a** *and* **Figure 2b** *as a function of $\|\zeta\|^2$, or equivalently as a function of*

$$t = \frac{\|\zeta\|^2}{d-1} = \frac{1}{d-1} \left\| \theta - \frac{a^\top \theta}{a^\top a} a \right\|^2 ,$$

i.e., in terms of the average squared component of ζ. The actual risks depend on the underlying mixing distributions $\mathcal{L}_1$ and $\mathcal{L}_2$, but not the risk differences as previously observed in Remark 4.1. Observe as well that t is independent of $\|a\|$ and only depends on the direction $a/\|a\|$. **Figure 2a** *has $\sigma_X^2 = 1, \sigma_Y^2 = 2$ and varying d, while* **Figure 2b** *has fixed $d = 5, \sigma_X^2 = 1$ with $\sigma_Y^2 = c\sigma_X^2$ and varying c. As seen with* **Figure 2a**, *the improvement in KL risk vanishes at $t \to \infty$, but gains in prominence with increasing d, and with the proximity of θ to the linear subspace spanned by a. As seen with* **Figure 2b**, *the KL risk difference loses in prominence with larger c which is consistent with the fact that MRE density gains in reliability when the variance σ_X^2 of the observable decreases.*

Frequentist risk ratios between $\hat{q}_U$ and $\hat{q}_{\pi_H}$ are plotted in **Figure 2c** *for $\sigma_X^2 = 1, \sigma_Y^2 = 2$ and varying d. These ratios depend additionally on the mixing distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ and they are set here with $\sqrt{\chi_1^2}$ mixing (Example 2.1 (B)), i.e., $X|\theta$ and $Y|\theta$ have skew-normal distributions with densities given in (2.3), and $\hat{q}_U$ given in part (D) of Section 2.2. We further set $a = \mathbf{1}_d = (1, \ldots, 1)^\top$, in which case the harmonic prior density on θ in (5.1) reduces to $\pi_H(\theta) = ||\theta - \bar{\theta}\mathbf{1}_d||^{-(d-3)}$ with $\bar{\theta} = \frac{1}{d}\sum_{i=1}^{d} \theta_i$. With the above settings, the constant (and minimax) risk of the MRE density can be computed from (2.10). For instance, we obtain $R(\theta, \hat{q}_U) \approx 1.0954$ for $d = 5$, ≈ 1.5187 for $d = 7$ and ≈ 1.9403 for $d = 9$. These are close to linear with the term $\frac{d}{2}\log\frac{\sigma_S^2}{\sigma_Y^2} = \frac{d}{2}\log\frac{3}{2}$ (≈ 1.0137 for $d = 5$, ≈ 1.4191 for $d = 7$ and ≈ 1.8246 for $d = 9$), representing the MRE risk for the normal case with $a = 0$, being dominant in (2.10). As seen in* **Figure 2c**, *where the risk ratios are plotted with respect to $t = \frac{1}{d-1}||\theta - \bar{\theta}\mathbf{1}_d||^2$, the gains increase in d and with the closeness of the θ_i's to $\bar{\theta}$.*

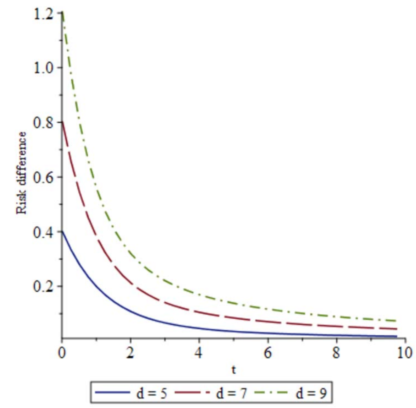
Example 5.2 (Plug-in type improved predictive density)**.** *In the context of Theorem 4.1 and Corollary 4.1), consider plug-in type predictive densities*

$$q_{\hat{\theta}_{JS}}(y; X) \sim MMN_d\left(\hat{\theta}_{JS}(X), a, \Sigma_X + \Sigma_Y, \mathcal{L}_3\right)$$

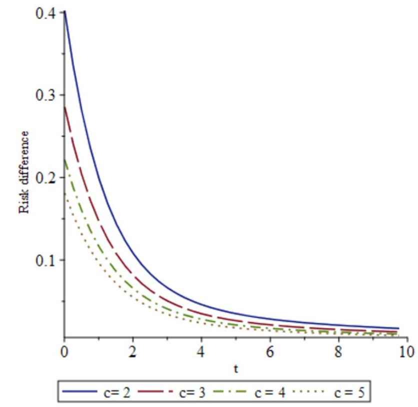
with the choice of the James-Stein estimator ($r(t) = d - 3$ in Corollary 4.1)

$$\hat{\theta}_{JS}(X) = X - \frac{(d-3)(1+c)}{\|H_2\Sigma_X^{-1/2}X\|^2}\Sigma_X^{1/2}H_2^\top H_2\Sigma_X^{-1/2}X, \tag{5.4}$$

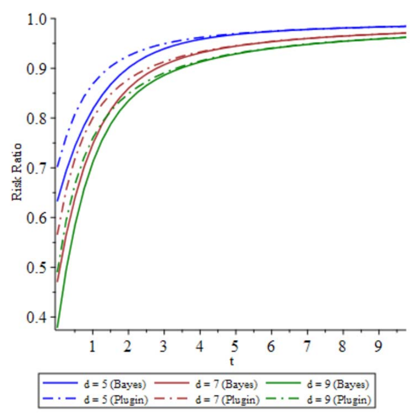
leading to the dominance of $q_{\hat{\theta}_{JS}}$ over $\hat{q}_U$ for $d \geq 4$. Both the dominating predictive density $q_{\hat{\theta}_{JS}}$ and the actual difference in risks do depend on the choice of H_2, but the KL risk difference, as given in (4.1) and mentioned at the end of Section 4.1, is independent of the underlying mixing distributions and will thus coincide with the corresponding difference stemming for $d-1$ dimensional normal models and which have appeared many times in the literature. The difference in risks will be a function of $\eta = H_2\Sigma_X^{-1/2}\theta$ in general, and more precisely as a function of $\|\eta\|^2$ in this case given that the James-Stein estimator is equivariant with respect to orthogonal transformations.



(a) The KL Risk difference between $\hat{q}_U$ and $\hat{q}_{\pi_H}$ as a function of $t = \frac{\|\zeta\|^2}{d-1} = \frac{1}{d-1}\left\|\theta - \frac{a^\top\theta}{a^\top a}a\right\|^2$, for $\sigma_X^2 = 1, \sigma_Y^2 = 2$.



(b) The KL Risk difference between $\hat{q}_U$ and $\hat{q}_{\pi_H}$ as a function of $t = \frac{\|\zeta\|^2}{d-1} = \frac{1}{d-1}\left\|\theta - \frac{a^\top\theta}{a^\top a}a\right\|^2$, for $d = 5$, $\sigma_X^2 = 1$, and $c = \frac{\sigma_Y^2}{\sigma_X^2} = 2, 3, 4, 5$.



(c) Kullback-Leibler risk ratios with: (i) $\hat{q}_U$ vs. q_{π_H} and (ii) $\hat{q}_U$ vs. $q_{\hat{\theta}_{JS}}$ as a function of $t = \frac{1}{d-1}\,||\theta - \bar{\theta}\mathbf{1}_d||^2$, for $\sigma_X^2 = 1, \sigma_Y^2 = 2$, where $\hat{\zeta}_{JS}$ is James-Stein estimator.

FIG 2. *KL risk performance of predictive density estimators.*

It is thus more interesting to look at the ratio of Kullback-Leibler risks and such ratios are presented in **Figure 2c** *with the same settings as in Example 5.1, i.e., multivariate skew-normal models with* $\sqrt{\chi_1^2}$ *mixing,* $\Sigma_X = \sigma_X^2 I_d = I_d, \Sigma_Y = \sigma_Y^2 I_d = 2I_d$*, and* $a = (1, \cdots, 1)^T$*. Again here, the risk ratios are plotted with respect to* $t = \frac{1}{d-1}\,||\theta - \bar{\theta}\mathbf{1}_d||^2$*, the gains increase in* d *and with the closeness of* *the θ_i's to $\bar{\theta}$.*

Remark 5.1. *As a follow-up to the examples above, we conclude by comparing the Bayes predictive density q_{π_H} and the plug-in type density $q_{\hat{\theta}_{JS}}$. While plug-in or plug-in type densities are simple and appealing to a broad spectrum of researchers, from a decision-theoretic perspective, it is often preferable to consider Bayes minimax procedure such as q_{π_H} and it is certainly of interest to compare the frequentist risk performance with other procedures. For the given settings (i.e., $d = 5, 7, 9$, $\sigma_X^2 = 1, \sigma_Y^2 = 2$),* **Figure 2c** *reveals better performance uniformly of q_{π_H}, and the same was observed for some other combinations of σ_X^2 and σ_Y^2 which we tested.*

Interestingly, as seen above with (4.1) and Remark 4.1, the difference in KL risks here will not depend on the settings of the perturbation vector a, and the mixing distributions $\mathcal{L}_1$ and $\mathcal{L}_2$. Therefore, the analysis for the normal case is critical. For the estimation of a multivariate normal mean θ under squared-error loss with covariance matrix $\sigma_X^2 I_p$, it was actually shown (i.e., [19]) that the Bayes estimator $\hat{\theta}_{\pi_H}$ with respect to the harmonic prior dominates the James-Stein estimator for $p \geq 3$. Given this finding, the historical parallels between the above mentioned point estimation problem and the normal model prediction problem, and the numerical evidence collected here, it certainly is worth further investigating whether there is a theoretical finding and justification of the dominance of q_{π_H} over $q_{\hat{\theta}_{JS}}$.

Concluding remarks

In this work, we have addressed the problem of determining efficient predictive densities under Kullback-Leibler frequentist risk for multivariate skew-normal distributions and, more generally, for mean-mixtures of multivariate normal (MMN) distributions, and provided Bayesian and plug-in type predictive densities which dominate the MRE density, and are minimax in four dimensions or more. In doing so, we have made use of a canonical transformation which leads to the decomposition of the Kullback-Leibler risk for the predictive densities being considered into two additive parts, one of which matching that of the MRE and minimax density, the other relating to a normal model and permitting improvement in view of shrinkage predictive density estimation results for such models. Further implications are provided for certain type of parametric restrictions. In addition, motivated by the relative paucity of analytical representations for Bayesian posterior and predictive densities, we have contributed such explicit representations.

This work represents, to the best of our knowledge, a first foray of the study of predictive density estimation for MMN distributions. The findings are thus novel and they are also unified. The canonical transformation technique may well find further applications in predictive analysis, such as for mean-variance mixture of normal distributions. Extensions to other choices of loss (e.g., α-divergence) and to unknown covariance structures would be most interesting to investigate as well. Finally, it would be particularly interesting to investigate analysis in the case of an i.i.d. draws from $X|\theta \sim MMN_d(\theta, a, \Sigma, \mathcal{L}_1)$ as the reduction for instance to a linear combination of the sample values (as described in Remark 2.1) is not a sufficiency reduction.

6. Appendix

A general representation of MMN distributions

This section contains further examples of MMN distributions as defined in Definition 1.1 based on the following lemma.

Lemma 6.1. *For a mixing density of the form*

$$\ell(v) = h(v)\, e^{-c_2 v - c_1 v^2/2}\, \mathbf{I}_{(0,\infty)}(v)\,, \tag{6.1}$$

with $c_1 > 0, c_2 \in \mathbf{R}$ or $c_1 = 0, c_2 \geq 0$, the corresponding pdf of X in (1.1) is given by

$$p(x|\theta) = \frac{1}{c_1'}\, \phi_d\,(x-\theta, \Sigma)\, \frac{\mathbf{E}\left[h\left\{ \frac{1}{c_1'}\left(Z + \frac{c_2'}{c_1'} \right) \right\} \Big|\, Z + \frac{c_2'}{c_1'} \geq 0 \right]}{R\left(\frac{c_2'}{c_1'} \right)}, \tag{6.2}$$

with $Z \sim N(0,1)$, $c_1' = \left(c_1 + a^\top \Sigma^{-1} a \right)^{1/2}$, $c_2' = (x-\theta)^\top \Sigma^{-1} a - c_2$, and $R(\cdot)$ the reverse Mill's ratio.

Proof. The result follows from (2.2) as the skewing factor $\mathbf{E}^V \left(e^{-\frac{V^2}{2} a^\top \Sigma^{-1} a} e^{V (x-\theta)^T \Sigma^{-1} a} \right)$ reduces to

$$\begin{aligned}
&\int_0^\infty e^{-\frac{v^2}{2}(c_1')^2} e^{v c_2'}\, h(v)\, dv \\
&= \frac{\sqrt{2\pi}}{c_1'}\, e^{\frac{c_2'^2}{2 c_1'^2}} \int_0^\infty h(v)\, \frac{c_1'}{\sqrt{2\pi}}\, e^{\frac{-c_1'^2 \left(v - \frac{c_2'}{(c_1')^2} \right)^2}{2}}\, dv \\
&= \frac{\sqrt{2\pi}}{c_1'}\, e^{\frac{c_2'^2}{2 c_1'^2}}\, \mathbf{E}\left\{ h\left(\frac{Z}{c_1'} + \frac{c_2'}{c_1'^2} \right) \Big|\, Z + \frac{c_2'}{c_1'} \geq 0 \right\} \Phi\left(\frac{c_2'}{c_1'} \right).
\end{aligned}$$

□

We point out that the above Lemma applies for $c_1 = c_2 = 0$ and thus covers all absolutely continuous distributions on $\mathbf{R}_+$. Here are nevertheless specific examples of model density (6.2).

Example 6.1. **(A)** *Gamma mixing with $\ell(v) = \frac{v^{\alpha-1} e^{-v/\beta}}{\Gamma(\alpha)\beta^\alpha}$. Lemma 6.1 applies with $h(v) = \frac{v^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}$, $c_1 = 0$ and $c_2 = 1/\beta$, and the model density is given by (6.2) with $c_1' = \left(a^\top \Sigma^{-1} a \right)^{1/2}$ and $c_2' = (x-\theta)^\top \Sigma^{-1} a - (1/\beta)$.*

The density was studied in [1, 2]. The exponential case with $\alpha = 1$ simplifies with

$$p(x|\theta) = \frac{1}{\beta c_1'} \; \frac{\phi_d\,(x-\theta;\Sigma)}{R\left(\frac{c_2'}{c_1'}\right)}\,. \tag{6.3}$$

More generally for positive integer α, the density's expression brings into play the $(\alpha-1)^{th}$ lower-truncated moment of a normal distribution. For instance, with $\mathbf{E}\,\{(Z+\Delta)|Z+\Delta \geq 0\} \;=\; \Delta \,+\, R(\Delta)$, we obtain for the case $\alpha = 2$ the model density:

$$p(x|\theta) \;=\; \frac{\phi_d(x-\theta,\Sigma)}{(c_1'\,\beta)^2}\,\left\{\frac{c_2'/c_1'}{R(c_2'/c_1')} \,+\, 1\right\}\,,$$

with the above c_1' and c_2'.

(B) $\sqrt{\chi^2_k}$ *mixing with* $h(v) \;=\; \frac{(\frac{1}{2})^{k/2-1}}{\Gamma(k/2)}\,v^{k-1}$, $c_1 \,=\, 1$, $c_2 \,=\, 0$, *and* $k \,>\, 0$. *The corresponding model density is given by (6.2) with the above h, $c_1' = \left(1+a^\top\Sigma^{-1}a\right)^{1/2}$, and $c_2' \;=\; (x-\theta)^\top\Sigma^{-1}a$. The density was given in [5] and, as previously noted, the case $k=1$ reduces to the skew-normal case in (2.3). As in Example* **(A)** *for positive integer k, the density's expression involves a lower-truncated moment of a normal distribution.*

(C) *Kummer type II mixing with $c_2 = c/\sigma$, $c_1 = 0$, $h(v) = \frac{\sigma^b}{\Gamma(a)\,\psi(a,1-b,c)}\,\frac{v^{a-1}}{(v+\sigma)^{a+b}}$ with $a,c,\sigma>0$, $b\in\mathbf{R}$, and ψ the confluent hypergeometric function of type II defined for $\gamma_1,\gamma_3>0$ and $\gamma_2\in\mathbf{R}$ as $\psi(\gamma_1,\gamma_2,\gamma_3) \;=\; \frac{1}{\Gamma(\gamma_1)}\int_{\mathbf{R}_+} t^{\gamma_1-1}(1+t)^{\gamma_2-\gamma_1-1}\,e^{-\gamma_3 t}\,dt$. This class of densities includes for $b=-a$ the Gamma densities in* **(A)***, as well as Beta type II densities for $c=0$ and $b>0$. The resulting mean-mixture density is given by (6.2) and involves interesting expectations of the form $\mathbf{E}\left(\frac{W^{a-1}}{(W+\sigma)^{a+b}}|W\geq 0\right)$ where $W \sim N(\Delta,1)$ with $\Delta = c_2'/c_1'$.*

Proof of Lemma 2.2. We have

$$\begin{aligned} e^{\frac{-AB^2}{2}}\,(2\pi A)^{-1/2}\int_0^\infty \Phi(ct)\,e^{-\frac{t^2}{2A}+Bt}\,dt \quad &= \quad \int_0^\infty \Phi(ct)\,\frac{1}{\sqrt{A}}\,\phi(\frac{t-AB}{\sqrt{A}})\,dt \\ &= \quad \mathbf{P}\,(U-cT\leq 0, -T\leq 0)\;, \end{aligned}$$

with U,T independently distributed as $N(0,1)$ and $N(\theta_T \,=\, AB, \sigma_T^2 \,=\, A)$, respectively. The result follows since

$$(U-cT,-T)^\top \sim N_2\left(\left(\begin{array}{r} -cAB \\ AB \end{array}\right), \left[\begin{array}{rr} 1+c^2A & cA \\ cA & A \end{array}\right]\right)\,. \qquad \square$$

Remark 6.1. *Predictive density estimates are intrinsic by nature which implies that the developments of Section 4.2, presented for $\Sigma_X = \sigma_X^2 I_d$ and $\Sigma_Y = \sigma_Y^2 I_d$ in model (1.1) with known σ_X^2 and σ_Y^2, apply as well for $\Sigma_Y = c\Sigma_X$ with known Σ_X, Σ_Y, and $c = \sigma_Y^2/\sigma_X^2$. Indeed, one can consider $X' = \Sigma_X^{-1/2}X$ for which $X|\theta \sim MMN_d(\Sigma_X^{-1/2}\theta, \Sigma_X^{-1/2}a, I_d, \mathcal{L}_1)$ to estimate the density of $Y' = \Sigma_X^{-1/2}Y$,* *for which $Y'|\theta \sim MMN_d(\Sigma_X^{-1/2}\theta, \Sigma_X^{-1/2}a, cI_d, \mathcal{L}_2)$. In doing so, one produces a predictive density estimator $q_1(y') = \hat{q}(y'; x'), y' \in \mathbf{R}^d$, for the density $q_{Y'}$ of Y', which equates to $q_2(y) = \hat{q}(\Sigma_X^{-1/2}y; \Sigma_X^{-1/2}x)\,|\Sigma_X^{-1/2}|; y \in \mathbf{R}^d$; as a predictive density estimator of the density q_Y of Y. Moreover, the Kullback-Leibler losses $\rho(q_{Y'}, q_1)$ and $\rho(q_Y, q_2)$ are equal, i.e.*

$$\int_{\mathbf{R}^d} q_{Y'}(t)\,\log\frac{q_{Y'}(t)}{q_1(t)}\,dt \;=\; \int_{\mathbf{R}^d} q_Y(t)\,\log\frac{q_Y(t)}{q_2(t)}\,dt\,,$$

as seen with the change of variables $t \to \Sigma_X^{-1/2}t$.

Proof of (5.2). With the standard representation $T|K \sim \chi^2_{d-1+2K}$ with $K \sim$ *Poisson* $\left(\frac{||z||^2}{2\sigma^2}\right)$, we have

$$\begin{aligned}
\mathbf{E}\,T^{(3-d)/2} &= \sum_{k=0}^{\infty} e^{-\frac{||z||^2}{2\sigma^2}}\,\frac{1}{k!}\left(\frac{||z||^2}{2\sigma^2}\right)^k \mathbf{E}\left(\chi^2_{d-1+2k}\right)^{\frac{(3-d)}{2}} \\
&= \frac{1}{2^{\frac{d-3}{2}}}e^{-\frac{||z||^2}{2\sigma^2}}\sum_{k=0}^{\infty}\left(\frac{||z||^2}{2\sigma^2}\right)^k \frac{1}{\Gamma(\frac{d-1}{2}+k)} \\
&= e^{-\frac{||z||^2}{2\sigma^2}}\left(\frac{||z||^2}{\sigma^2}\right)^{-\frac{d-3}{2}} \sum_{k=\frac{d-3}{2}}^{\infty}\left(\frac{||z||^2}{2\sigma^2}\right)^k \frac{1}{k!}\,,
\end{aligned}$$

which yields (5.2). □

Acknowledgments

Thanks to Jean-Philippe Burelle for useful discussions on geometric representations related to prior density (4.9). Finally, thanks to two reviewers, an Associate Editor, and Editor Grace Yi for valuable comments and suggestions throughout the evaluation process.

References

[1] Abdi, M., Madadi, M., Balakrishnan, N. & Jamalizadeh, A. (2021). Family of mean-mixtures of multivariate normal distributions: Properties, inference and assessment of multivariate skewness *Journal of Multivariate Analysis*, **181**, 104679. MR4161850

[2] Adcock, C.J. & Shutes, K. (2012). On the multivariate extended skew-normal, normal-exponential and normal-gamma distributions. *Journal of Statistical Theory and Practice*, **6**, 636–664. MR3196571

[3] Aitchison, J. & Dunsmore, I.R. (1975). Statistical Prediction Analysis. *Cambridge University Press.* MR0408097

[4] Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554. MR0391353

[5] Arrellano-Valle, R.B. & Azzalini, A. (2021). A formulation for continuous mixtures of multivariate normal distributions. *Journal of Multivariate Analysis*, **185**, 104780. MR4277717

[6] Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, **83**, 715–726. MR1440039

[7] Barndorff-Nielsen, O., Kent, J. & Sørensen, O. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review*, **50**, 145–159. MR0678296

[8] Bock, M.E., Judge, G.G. & Yancey, T.A. (1984). A simple form for the inverse moments of a non-central χ^2 and F random variables and certain confluent hypergeometric functions. *Journal of Econometrics*, **25**, 217–234. MR0748045

[9] Contreras-Reyes, J.E. & Arellano-Valle, R.B. (2012). Kullback-Leibler divergence measure for multivariate skew-normal distributions. *Entropy*, **14**, 1606–1626. MR2979804

[10] Fourdrinier, D., Marchand, É., Righi, A., & Strawderman, W.E. (2011). On improved predictive density estimation with parametric constraints. *Electronic Journal of Statistics*, **5**, 172–191. MR2792550

[11] Fourdrinier, D., Strawderman, W.E. & Wells, M.T. (2018). Shrinkage estimation. Springer series in statistics. Springer. New York, Dordrecht, Heidelberg, London.

[12] George, E., Marchand, É., Mukherjee, G. & Paul, D. (2019). New and evolving roles of shrinkage in large-scale prediction and inference. BIRS Workshop Report.

[13] George, E.I., Liang, F. & Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Annals of Statistics*, **34**, 78–91. MR2275235

[14] Komaki, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika*, **88**, 859–864. MR1859415

[15] Kubokawa, T., Marchand, É. & Strawderman, W.E. (2015). On predictive density estimation for location families under integrated squared error loss. *Journal of Multivariate Analysis*, **142**, 57–74. MR3412739

[16] Kubokawa, T., Marchand, É., Strawderman, W.E. & Turcotte, J.P. (2013). Minimaxity in predictive density estimation with parametric constraints. *Journal of Multivariate Analysis*, **116**, 382–397. MR3049911

[17] Kubokawa, T., Strawderman, W.E. & Yuasa, R. (2020). Shrinkage estimation of location parameters in a multivariate skew-normal distribution. *Communications in Statistics: Theory and Methods*, **49**, 2008–2024. MR4070897

[18] Kubokawa, T. (2005) Estimation of bounded location and scale parameters. *Journal of the Japanese Statistical Society*, **35**, 221–249. MR2328426

[19] Kubokawa, T. (1991) Improving the James-Stein estimator. *Journal of Multivariate Analysis*, **36**, 121–126. MR1094273

[20] Liang, F. & Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, **50**, 2708–2726 MR2096988

[21] LMoudden, A. & Marchand, É. (2021). Bayesian estimation and prediction for certain type of mixtures. *Communications in Statistics: Theory and Methods*, DOI:10.1080/03610926.2021. 1913185.

[22] Marchand, É. & Sadeghkhani, N. (2018). On predictive density estimation with additional information. *Electronic Journal of Statistics*, **12**, 4209–4238. MR3892140

[23] Marchand, É. & Strawderman, W.E. (2004). Estimation in Restricted Parameter Spaces: A Review, *Festschrift for Herman Rubin*. Institute of Mathematical Statistics Lecture Notes-Monograph Series, pp. 21–44. MR2126884

[24] Park, J. (2021). Finding Bayesian optimal portfolios with skew normal returns. Available at SSRN 3640062.

[25] Negarestanu, H., Jamalizadeh, A., Shafiei, S. & Balakrishnan, N. (2019). Mean mixtures of normal distributions: Properties, inference and applications. *Metrika*, **82**, 501–528. MR3947858

[26] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, I, pp. 197–206. University of California Press. MR0084922

[27] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135–1151. MR0630098

[28] van Eeden, C. (2006). *Restricted parameter space problems: Admissibility and minimaxity properties*. Lecture Notes in Statistics, **188**, Springer.