Using the compensation identities to analyze the risk of density estimators

W. D. Brinda and Joseph T. Chang

The *compensation identity* states that the expected I-divergence[1] from a random distribution to a fixed distribution equals the expected I-divergence from the random distribution to its centroid plus the I-divergence from the centroid distribution to the fixed distribution. The *reverse compensation identity* has the fixed and random distributions in the other order and uses a differently-defined centroid; it states that the expected I-divergence from a fixed distribution to a random distribution equals the expected I-divergence from a centroid to the random distribution plus the I-divergence from the fixed distribution to that centroid. These two identities are information-theoretic analogues of the bias-variance decomposition, and as such they provide decompositions that can be enlightening when analyzing the risk of density estimators.

In Section 1, we formally define the compensation identities and explain how they decompose the I-divergence risk of density estimators. Next, Section 2 describes variational Bayesian estimators which we will later use to demonstrate the decompositions. Incidentally, the mean field algorithm used in variational Bayesian procedures can be easily understood in light of one of the compensation identities, an observation that we will highlight in our discussion. In Section 3, the reverse compensation decomposition is worked out for a variety of Gaussian location estimators, and their bias-like and variance-like terms are compared. Finally, Section 4 uses the context of Gaussian mixture estimation to demonstrate how simulations can be used to understand the compensation decomposition when the quantities involved are analytically intractable.

1 The compensation identities

Theorem 1.1, called the *compensation identity* by [Topsøe, 2001, Thm 9.1], conveniently decomposes the expected I-divergence from a random probability measure to a fixed probability measure.[2]

Theorem 1.1 (The compensation identity)**.** *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure μ, and suppose that*

[1] I divergence stands for information divergence; it is more commonly known as Kullback divergence or relative entropy.

[2] In Theorem 1.1 and throughout the remainder of this paper, lower-case and upper-case letters implicitly pair probability measures with their densities.

$(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be an $\mathcal{X}$-valued random element. For any probability measure R on $\mathcal{Y}$,

$$\mathbb{E}D(Q_X\|R) = D(\overline{Q}_P\|R) + \mathbb{E}D(Q_X\|\overline{Q}_P)$$

where $\overline{Q}_P$ represents the P-mixture over $\{q_x\}$.

A less familiar decomposition, which we will call the *reverse compensation identity*, holds when the expected I-divergence's *second* argument is random rather than its first. Instead of a mixture, it involves a *geometric mixture*.[3] We define the *P-geometric mixture* of $\{q_x\}$ to be the probability measure with density

$$\widetilde{Q}_P(y) := \frac{e^{\mathbb{E}_{X\sim P} \log q_X(y)}}{\int e^{\mathbb{E}_{X\sim P} \log q_X(y)} d\mu(y)}.$$

Jensen's inequality and Tonelli's theorem together provide an upper bound for the denominator.

$$\begin{aligned}\int e^{\mathbb{E} \log q_X(y)} d\mu(y) &\leq \mathbb{E} \int e^{\log q_X(y)} d\mu(y) \\ &= 1\end{aligned}$$

This integral can be zero, however, in which case the geometric mixture is not well-defined.[4]

Theorem 1.2 (The reverse compensation identity)**.** *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure μ, and suppose that $(x, y) \mapsto q_x(y)$ is product measurable. Let $X \sim P$ be an $\mathcal{X}$-valued random element. If $\int e^{\mathbb{E} \log q_X(y)} d\mu(y) > 0$, then for any probability measure R on $\mathcal{Y}$,*

$$\mathbb{E}D(R\|Q_X) = D(R\|\widetilde{Q}_P) + \mathbb{E}D(\widetilde{Q}_P\|Q_X)$$

where $\widetilde{Q}_P$ represents the P-geometric mixture over $\{q_x\}$.

Proofs of Theorems 1.1 and 1.2 can be found in [Brinda, 2018, Appendix A].

Theorems 1.1 and 1.2 are perfectly analogous to the bias-variance decomposition for Hilbert-space-valued random vectors.[5] The expected divergence from the a random element to a fixed element decomposes into the divergence from a "centroid" of the random element to that fixed element plus the internal variation of the random element from that centroid.[6] We suggest a notation that

[3] What we call a "geometric mixture" is sometimes called a "log mixture" or "log-convex mixture," for instance by [Grünwald, 2007, Sec 19.6].

[4] An example of such a pathological case is when q_X has positive probabilities on two densities that are mutually singular.

[5] In fact, the compensation identity and bias-variance decomposition are both instances of this decomposition for Bregman divergences — see [Telgarsky and Dasgupta, 2012, Lem 3.5] and Pfau [2013].

[6] It follows that the centroid is the choice of fixed element that has the smallest possible expected divergence from the random element.

makes use of this intuition:

$$\begin{aligned}\bar{\mathbb{V}}Q_X :&= \inf_R \mathbb{E}D(Q_X\|R)\\ &= \mathbb{E}D(Q_X\|\overline{Q}_P)\end{aligned}$$

and[7]

$$\begin{aligned}\widetilde{\mathbb{V}}Q_X :&= \inf_R \mathbb{E}D(R\|Q_X)\\ &= \begin{cases}\mathbb{E}D(\widetilde{Q}_P\|Q_X), & \text{if } \int e^{\mathbb{E}\log q_X(y)}d\mu(y) > 0\\ \infty, & \text{otherwise.}\end{cases}\end{aligned}$$

We also suggest the terminology *information risk* (*I-risk*), *information bias* (*I-bias*) squared, and *information variance* (*I-variance*) for the quantities in the compensation identity as well as the terminology *reverse information risk* (*rI-risk*), *reverse information bias* (*rI-bias*) squared, and *reverse information variance* (*rI-variance*) for the quantities in the reverse compensation identity.[8] The language introduced here comports with that of information projections (I-projections) and reverse information projections (rI-projections).

The compensation identities can provide insights regarding regularization, and we conclude this subsection with one such observation. A simple way to regularize a point-estimator $\hat{\theta}$ is by shrinking it toward any constant point θ_0. The variance of $[1-\lambda]\hat{\theta} + \lambda\theta_0$ is $[1-\lambda]^2$ times the variance of the original estimator $\hat{\theta}$. Similarly, a density estimator's I-variance can always be decreased by *mixing with a fixed distribution.*

Theorem 1.3. *Let $\{q_x : x \in \mathcal{X}\}$ be a family of probability densities with respect to a σ-finite measure γ, and suppose that $(x,y) \mapsto q_x(y)$ is product measurable. Let X be an $\mathcal{X}$-valued random element. For any fixed known probability measure $\check{Q}$, the I-variance of the mixture $\bar{\mathbb{V}}([1-\lambda]Q_X + \lambda\check{Q})$ is non-increasing as $\lambda \in [0,1]$ increases. The I-variance is strictly decreasing unless Q_X equals $\check{Q}$ with probability one.*

Proof. With P representing the distribution of X, the mixture's centroid is $[1-\lambda]\overline{Q}_P + \lambda\check{Q}$.

For any $\lambda_1 \in [\lambda, 1]$, a draw from $[1-\lambda_1]Q_x + \lambda_1\check{Q}$ can be achieved by "processing" a draw from $[1-\lambda]Q_x + \lambda\check{Q}$. One simply needs to switch it to a new draw from $\check{Q}$ with probability $\frac{\lambda_1-\lambda}{1-\lambda}$. The data processing inequality tells us that two processed distributions are no further in relative entropy than the unprocessed distributions were.

The same processing that transforms $[1-\lambda_1]Q_x+\lambda_1\check{Q}$ to $[1-\lambda]Q_x+\lambda\check{Q}$ also transforms the centroids appropriately. Thus by the data processing inequality,

$$D([1-\lambda_1]Q_x + \lambda_1\check{Q}\,\|\,[1-\lambda_1]\overline{Q}_P + \lambda_1\check{Q}) \le D([1-\lambda]Q_x + \lambda\check{Q}\,\|\,[1-\lambda]\overline{Q}_P + \lambda\check{Q}).$$

Since this holds for every $x \in \mathcal{X}$, it holds for any expectation over $\mathcal{X}$. □

[7] This alternative representation of $\widetilde{\mathbb{V}}$ is justified by [Brinda, 2018, Lem A.3.4].

[8] For information-theoretic interpretations of these quantities, see [Brinda, 2018, Sec A.1].

USE an abstract Taylor expansion to show that the I-variance behaves proportionally to $(1-\lambda)^2$ plus lower order terms. What type of abstract derivatives are needed? Frechet? Make sure to build in enough conditions to guarantee that needed derivatives exist - then if it's challenging to state a closed-form expression just use a symbol e.g. "$\frac{(1-\lambda)^2}{2}\mathbb{E}s_X \cdot (Q_X - \overline{Q}_P)^2 + O([1-\lambda]^3)$ where s is the second Frechet(?) derivative - how can I guarantee that it's finite? it wouldn't be the end of the world to just add the condition that $\mathbb{E}s_X \cdot (Q_X - \overline{Q}_P)^2$ is finite.

CHECK to see if the squared I-bias behaves proportionally to λ^2 plus lower order terms when Q_X is unbiased - if so, then state a corollary that for any P and $\check{Q}$ the risk of an estimator with zero I-bias can always be improved by regularizing with some positive λ. Presumably similar results hold for the rI-risk quantities. Move all proofs to the end.

2 Variational Bayesian estimators

Calculus of variations is the study of optimization over a space of functionals. *Variational approximation* means identifying the functional in a set that is *closest* to a fixed target functional. When the target functional is a probability measure with a density only known up to a constant, the task of identifying the closest probability measure in a set is *variational Bayesian inference*. I-divergence (with either order of arguments) is a commonly used divergence in this context. To begin this section, we provide a straight-forward explanation of the mean field algorithm by making reference to the reverse compensation identity. Then we point to *variational posterior mixtures* as density estimators of interest.

2.1 Mean field approximation

When I-divergence (with the target as the second argument) is used to quantify closeness and the search space comprises all probability measures with a specific product structure, the variational Bayes problem is called *mean field approximation*. The approximating distribution is the information projection of the target onto the set of all probability measures with the specified product structure. Inspired by the *mean field theory* of physics, Ghahramani [1995] introduced this technique for statistical learning.

Suppose some target distribution on $\mathcal{X} \times \mathcal{Y}$ can be represented as $P \otimes \{Q_x\}$ for some probability measure P on $\mathcal{X}$ and a probability kernel $\{Q_x : x \in \mathcal{X}\}$ of "conditional distributions" with densities $\{q_x\}$ relative to a σ-finite dominating measure. The relative entropy from any product measure $\check{P} \otimes \check{Q}$ to the target

is[9]

$$D(\check{P} \otimes \check{Q} \| P \otimes \{Q_x\}) = \mathbb{E}_{X\sim\check{P}} \mathbb{E}_{Y\sim\check{Q}} \log \frac{\check{p}(X)\check{q}(X)}{p(X)q_X(Y)}$$
$$= D(\check{P}\|P) + \mathbb{E}_{X\sim\check{P}} D(\check{Q}\|Q_X). \tag{1}$$

The reverse compensation identity (Theorem 1.2) implies that for any given $\check{P}$, the optimal choice of $\check{Q}$ is the $\check{P}$-geometric mixture of $\{Q_x\}$. Likewise, if the target distribution also has a representation as $Q \otimes \{P_y\}$ with the roles of marginal and conditional variable reversed, then the $\check{Q}$-geometric mixture of $\{P_y\}$ is the optimal choice of $\check{P}$ for fixed $\check{Q}$. The same logic continues to hold if the product structure has more than two components: any one component to be optimized plays the role of $\check{Q}$ in (1) while the rest of the components together play the role of $\check{P}$. The *mean field algorithm* constructs a product measure approximation by cycling through the components in this manner, updating each piece by setting it to the appropriate geometric mixture.[10] Equation (1) makes it clear that the algorithm is monotonic; each step can only decrease the relative entropy from the product approximation to the target; furthermore, it is guaranteed to converge to a local optimum [Bishop, 2006, Sec 10.1.1].

In Bayesian analysis, the posterior distribution represents an appropriate "belief" about the unknown parameter that arises from updating a prior belief based on observed data. However, posterior probabilities of parameter regions and posterior expectations of functions of the parameters are often challenging to calculate. If the parameters have a conjugate prior, then integrals can be calculated analytically; if the dimension of the parameter space is small, then integrals can be calculated numerically. Otherwise, practitioners turn to a variety of other approaches. Markov Chain Monte Carlo methods attempt to generate samples from the true posterior, but it is time-consuming and can do poorly when the posterior is badly multi-modal. Alternatively, the posterior's mean field approximation can have an analytically tractable form. Most convenient is when each conditional family is an exponential family, in which case the geometric mixture is itself in that family as well[11]; one simply needs to update the hyper-parameters to identify the new distribution.

Consider reversing the order of arguments; let us ask what is the best product approximation when the target distribution is the first argument of I-divergence.

$$D(P \otimes \{Q_x\} \| \check{P} \otimes \check{Q}) = D(P\|\check{P}) + \mathbb{E}_{X\sim P} D(Q_X\|\check{Q})$$

[9] The freedom to choose the order of integration is justified by Tonelli's theorem because there is a an alternative representation of relative entropy with a non-negative integrand. One source with a clear explanation is [Brinda and Klusowski, Submitted to Bernoulli in 2018, Lem 3.1].

[10] Most sources explaining the mean field algorithm put the joint distributions in place of the conditional distributions in the expression that we call the "geometric mixture." Both definitions result in the same distribution, so one can use whichever is more convenient.

[11] The $\check{P}$-geometric mixture over an exponential family is the distribution corresponding to the expectation of the canonical parameter, so the problem is simplest when x is indexing a canonical parameterization.

No matter what $\check{P}$ has been used, the optimal choice of $\check{Q}$ is the P-mixture of $\{Q_x\}$, according to the compensation identity (Theorem 1.1). Likewise, the optimal $\check{P}$ is the Q-mixture of $\{P_y\}$. These are precisely the marginals of the target distribution, and the resulting relative entropy to the approximation is the mutual information. The *expectation propagation algorithm* was devised by Minka [2001] to seek the marginal distributions when the normalizing factor is unknown. In general, the mean field approximation is more concentrated than the product of the marginals [Bishop, 2006, Sec 10.1.2].

Variational Bayesian methods may be useful for calculating approximate posteriors, but they are also "statistically unsound" in a sense. Ideally, a statistical procedure should be eventually correct, if enough data is collected and the algorithm runs long enough. However, if for instance the correct posterior belief is that the variables are highly correlated, the product approximations will never indicate that belief regardless of the amount of data and run-time. Any change in the scale of a probability measure will have an exactly corresponding change in the scale of the mean field approximation and the product of marginals approximation since relative entropy is scale-invariant. The resulting divergence between the probability measure and its approximation will remain unchanged in terms of relative entropy or any other f-divergence. Thus as a probability measure becomes more concentrated, these product approximations do not get closer to it, at least in terms of scale-invariant divergences. Variational Bayesian methods trade correctness for convenience, but their popularity is a sign that this trade-off is sometimes worth taking.

2.2 Variational posterior mixtures

In Bayesian analysis, the mixture over the model distributions using the posterior distribution for the mixing weights is called the *posterior mixture*. It represents the Bayesian's belief about what the next datum will be. In a variational Bayesian analysis, one might instead use the approximate posterior as mixing weights to calculate what we will call the *variational posterior mixture*, or more specifically either the *mean field mixture* or *product of marginals mixture*; this is demonstrated in [Bishop, 2006, Sec 10.2.3].

If the posterior becomes increasingly concentrated and the model is reasonably smooth, then both the posterior mixture and the variational posterior mixture will come to resemble the MAP distribution. In fact, one might think of the variational mixture as being *between* the posterior mixture and the MAP. In light of the compensation identities, we will explore whether this comparison has an interpretation in terms of the I-divergence risk analogues of bias and variance.

To work out the risk in the upcoming example, we will need to know the mean field approximation of a d-dimensional Gaussian distribution with mean θ and precision matrix Λ. As [Bishop, 2006, Sec 10.1.2] explains, each product component is Gaussian and inherits its means from the original distribution and

inherits its precision matrix[12] from the corresponding rows and columns of Λ.

3 Risk of Gaussian location estimators

Let $X^n := (X_1, \ldots, X_n) \overset{iid}{\sim} P$, and assume P has finite second moments, letting μ and Σ denote its mean and covariance matrix. We will compare a handful of density estimators based on a Gaussian location model $\{N(\theta, I_d) : \theta \in \mathbb{R}^d\}$. For any estimator Q_{X^n}, we can decompose the risk into "bias squared" and "variance" terms

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} D(P \| Q_{X^n}) = D(P \| \widetilde{Q}_{P^n}) + \mathbb{E}_{X^n \overset{iid}{\sim} P} D(\widetilde{Q}_{P^n} \| Q_{X^n});$$

these quantities are called the rI-risk, rI-bias squared, and rI-variance in Section 1.

If the estimator Q_{X^n} is Gaussian $N(\hat{\theta}(X^n), C)$ for some covariance C that does not depend on the data, then $\widetilde{Q}_{P^n}$ is $N(\widetilde{\theta}, C)$ where $\widetilde{\theta} := \mathbb{E}_{X^n \overset{iid}{\sim} P} \hat{\theta}(X^n)$, then the rI-bias squared simplifies conveniently as

$$\begin{aligned} D(P \| \widetilde{Q}_{P^n}) &= \tfrac{1}{2} \mathbb{E}_{Y \sim P} [Y - \widetilde{\theta}]' \Sigma^{-1} [Y - \widetilde{\theta}] + \tfrac{1}{2} \log[(2\pi)^d |C|] - h(P) \\ &= \tfrac{1}{2} \left(\operatorname{tr}(C^{-1}\Sigma) + [\mu - \widetilde{\theta}]' C^{-1} [\mu - \widetilde{\theta}] + \log |C| \right) + \underbrace{\tfrac{d}{2} \log(2\pi) - h(P)}_{\text{“}z_P\text{”}}. \end{aligned}$$

The rI-variance simplifies as well, if we interchange the order of integration (to justify the interchange, again see [Brinda, 2018, Lem A.3.1])

$$\begin{aligned} \mathbb{E}_{X^n \overset{iid}{\sim} P} D(\widetilde{Q}_{P^n} \| Q_{X^n}) &= \tfrac{1}{2} \mathbb{E}_{Y \sim P} \mathbb{E}_{X^n \overset{iid}{\sim} P} \left([Y - \hat{\theta}(X^n)]' C^{-1} [Y - \hat{\theta}(X^n)] - [Y - \widetilde{\theta}]' C^{-1} [Y - \widetilde{\theta}] \right) \\ &= \mathbb{E}_{X^n \overset{iid}{\sim} P} [\hat{\theta}(X^n) - \widetilde{\theta}]' C^{-1} [\hat{\theta}(X^n) - \widetilde{\theta}]. \end{aligned}$$

Assuming additionally that the location estimate is a linear transformation T of the sample mean, the rI-variance simplifies further. $\hat{\theta}(X^n) = T\overline{X}$ implies $\widetilde{\theta} := \mathbb{E}\hat{\theta}(X^n)$ is $T\mu$, so the rI-variance term is

$$\begin{aligned} \mathbb{E}_{X^n \overset{iid}{\sim} P} [T\overline{X} - T\mu]' C^{-1} [T\overline{X} - T\mu] &= \mathbb{E}_{X^n \overset{iid}{\sim} P} [\overline{X} - \mu]' T' C^{-1} T [\overline{X} - \mu] \\ &= \tfrac{1}{n} \operatorname{tr}(T' C^{-1} T \Sigma). \end{aligned}$$

Remarkably, *the risk only depends on P via μ and Σ*, except for the differential entropy term in z_P.

We will compare five estimators in this context: maximum likelihood, MAP, posterior mixture, mean field mixture, and product of marginals mixture.[13] We will see that all of them satisfy the assumptions posited above in that they map X^n to a normal location family and that the selected location $\hat{\theta}(X^n)$ is a linear transformation of the sample mean.

[12] In contrast, the product of marginals approximation inherits its covariance matrix from the original Gaussian distribution.

[13] The latter two were defined in .

The $\{N(\theta, I_d) : \theta \in \mathbb{R}^d\}$ model's maximum likelihood estimator gives the random distribution $Q_{X^n} = N(\overline{X}, I_d)$. For that estimator, Σ is I_d, $\hat{\theta}(X^n)$ is $\overline{X}$, $\widetilde{\theta}$ is μ, and the rI-bias squared and rI-variance simplify to

$$D(P\|\widetilde{Q}_{P^n}) = \tfrac{1}{2}\mathrm{tr}(\Sigma) + z_P$$

and

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} D(\widetilde{Q}_{P^n}\|Q_{X^n}) = \tfrac{1}{n}\mathrm{tr}(\Sigma)$$

As a sanity check, observe that if P is indeed Gaussian with identity covariance, the rI-bias is zero and the rI-variance is $\mathbb{E}\|X_1 - \mu\|^2/n$.

Next, we consider estimators that can arise from Bayesian analysis using a $N(0, V_0)$ prior on the location; let $\Lambda_0 := V_0^{-1}$ be the prior's precision. The posterior is also Gaussian with precision $\Lambda_n = \Lambda_0 + nI_d$ and location $nV_n\overline{X}$, where V_n denotes Λ_n^{-1}, the posterior's covariance — see [Murphy, 2007, Sec 7].

The MAP is the distribution $Q_{X^n} = N(nV_n\overline{X}, I_d)$. It has expected location $\widetilde{\theta} = nV_n\mu$, and its rI-risk decomposition has

$$D(P\|\widetilde{Q}_{P^n}) = \tfrac{1}{2}\mathrm{tr}(\Sigma) + \|\mu - nV_n\mu\|^2 + z_P$$

and

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} D(\widetilde{Q}_{P^n}\|Q_{X^n}) = n\,\mathrm{tr}(V_n' V_n \Sigma).$$

The posterior mixture is a Gaussian weighting over a Gaussian location model, which results in another Gaussian distribution. The mean of the resulting Gaussian is equal to the mean of the weighting distribution, in our case $nV_n\overline{X}$. The resulting Gaussian's covariance is larger than that of the model, however; it can be seen using [Bishop, 2006, Sec 2.3.3] that the resulting covariance is the sum of the mixing covariance and model covariance, in our case $V_n + I_d$.

The posterior mixture is therefore $Q_{X^n} = N(nV_n\overline{X}, V_n + I_d)$ and has rI-bias squared

$$D(P\|\widetilde{Q}_{P^n}) = \tfrac{1}{2}\mathrm{tr}((V_n + I_d)^{-1}\Sigma) + [\mu - nV_n\mu]'(V_n + I_d)^{-1}[\mu - nV_n\mu] + \tfrac{1}{2}\log|V_n + I_d| + z_P$$

and rI-variance

$$\mathbb{E}_{X^n \overset{iid}{\sim} P} D(\widetilde{Q}_{P^n}\|Q_{X^n}) = n\,\mathrm{tr}(V_n'(V_n + I_d)^{-1} V_n \Sigma). \tag{2}$$

Consider the mean field approximation that imposes mutual independence on all of the posterior's coordinates. Recall from our earlier discussion that the posterior's mean field approximation with independent coordinates has the same mean $nV_n\overline{X}$ and has precision $\mathrm{diag}(\Lambda_n)$. Therefore the mean field posterior mixtures is $Q_{X^n} = N(nV_n\overline{X}, [\mathrm{diag}(\Lambda_n)]^{-1} + I_d)$. Its rI-bias squared and rI-variance look like the posterior mixture's, except that instances of the posterior mixture's covariance $V_n + I_d$ should instead be the mean field approximation's covariance

$[\text{diag}(\Lambda_n)]^{-1} + I_d$. Similarly for the product-of-marginals approximation imposing independent coordinates, which is $Q_{X^n} = N(nV_n\overline{X}, \text{diag}(V_n) + I_d)$ so the posterior mixture's precision must be replaced with the precision $\text{diag}(V_n) + I_d$ in the rI-risk quantities.

Now we will visually compare the rI-biases, rI-variances, and total rI-risks of these estimators when the dimension is $d = 2$ and the sample size is $n = 1$. Suppose that P has expectation $\mu = \frac{r}{\sqrt{2}}(1, 1)'$ and covariance

$$\Sigma = s \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

For maximum likelihood, the rI-risk quantities (ignoring z_p) will not depend on a or ρ. The rI-bias squared is $s + z_P$ and the rI-variance is s/n for a total rI-risk of $(1 + \frac{1}{n})s + z_P$.

The behavior of the Bayesian estimators depend on ρ and a in addition to s. For zero ($\rho = 0$), low ($\rho = .25$), and high ($\rho = .75$) correlations respectively, Figures 1, 2, and 3 provide heat maps over the (a, s)-plane. These plots indicate the relative behavior of our estimators over a rather large portion of the space of possible data-generating distributions.[14]

If the chosen prior covariance is diagonal, then so is the posterior covariance. In that case the mean field and the product of marginals approximations are both the same as the actual posterior. To ensure that our plots highlight the differences that can result from using the approximate versus the true posterior, we will use a prior of

$$V_0 = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix}.$$

Equation (2) makes it clear that each estimator's rI-variance is actually linear in s and has no dependence on a. Figure 4 provides a better view and shows both variational Bayesian approximations between the MAP and the posterior mixture for the data-generating covariances under consideration.that figure is completely unnecessary. does this provide evidence against a regularizing effect of the variational Bayesian approximations? COMMENT???

In terms of both rI-bias and rI-variance, our visualizations show the variational Bayesian approximation mixtures behaving "in between" the MAP and posterior mixture. Furthermore, for some true distributions, the rI-risk of the variational Bayesian approximation can be lower than the rI-risk of the true Bayesian procedure being approximated. However, we note that the posterior mixture has smaller expected rI-risk if the averaging is taken with respect to the prior over the model — see [Brinda, 2018, Sec A.2].

[14] This claim of a "rather large portion" refers to the many symmetries inherent in the problem and to the fact that our analysis did not assume any particular form for the distribution P.

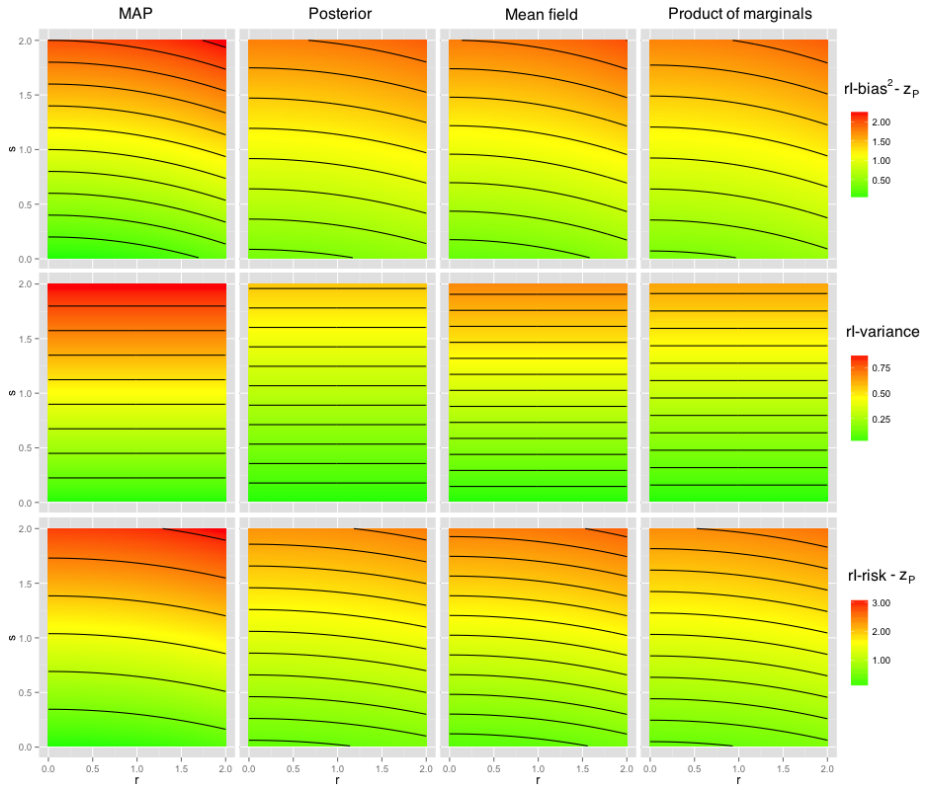

Figure 1: A sample of size is $n = 1$ is to be taken from a data-generating distribution with expectation $\frac{r}{\sqrt{2}}(1, 1)'$, marginal standard deviations s, and correlation $\rho = 0$. The heat maps show the reverse compensation identity's quantities for four Bayesian estimators of standard Gaussian location using a Gaussian prior with mean zero, both standard deviations 1, and correlation .75.

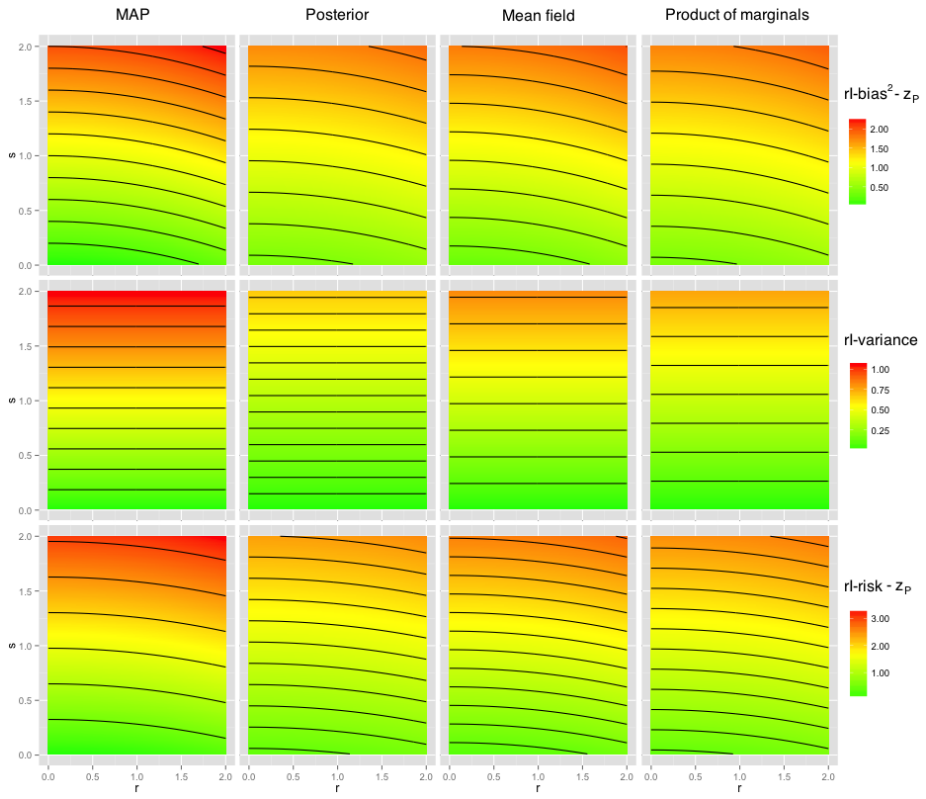

Figure 2: A sample of size is $n = 1$ is to be taken from a data-generating distribution with expectation $\frac{r}{\sqrt{2}}(1, 1)'$, marginal standard deviations s, and correlation $\rho = .25$. The heat maps show the reverse compensation identity's quantities for four Bayesian estimators of standard Gaussian location using a Gaussian prior with mean zero, both standard deviations 1, and correlation .75.

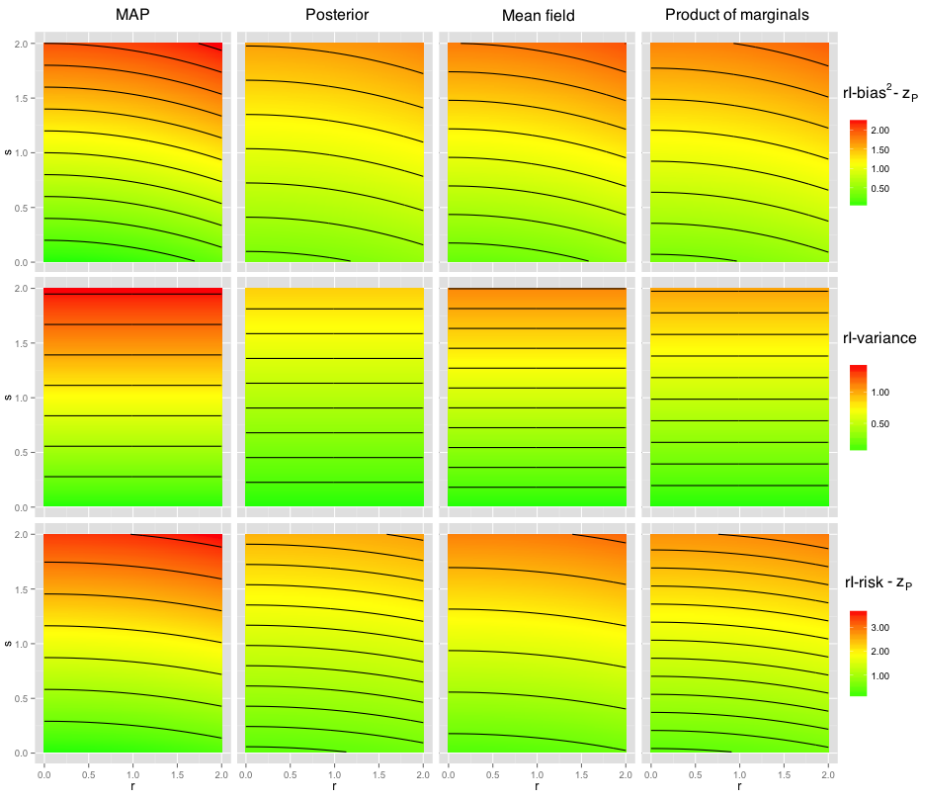

Figure 3: A sample of size is $n = 1$ is to be taken from a data-generating distribution with expectation $\frac{r}{\sqrt{2}}(1, 1)'$, marginal standard deviations s, and correlation $\rho = .75$. The heat maps show the reverse compensation identity's quantities for four Bayesian estimators of standard Gaussian location using a Gaussian prior with mean zero, both standard deviations 1, and correlation .75.

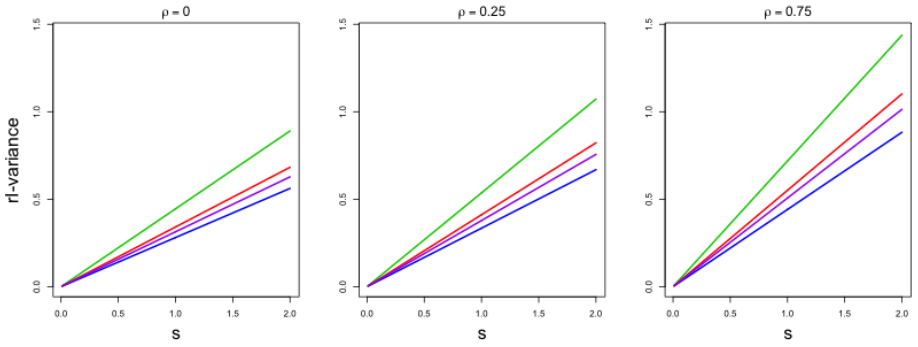

Figure 4: A sample of size is $n = 1$ is to be taken from a two-dimensional data-generating distribution with marginal standard deviations s and three possible correlations $\rho = 0$, .25, and .75. The lines show the rI-variances of four Bayesian estimators of standard Gaussian location using a Gaussian prior with mean zero, both standard deviations 1, and correlation .75. MAP is in green, posterior mixture is in blue, mean field mixture is in red, and product of marginals mixture is in purple.

4 Risk of Bayesian Gaussian mixture estimators

Next, we will consider how variational approximation affects risk in the context of Gaussian mixtures. Specifically, consider Bayesian inference with a prior that has independent standard Gaussians for the component means and has independent uniform multi-Bernoulli distributions for the labels. Furthermore, the prior has all component means independent of all labels. The true posterior can have dependence between the component means and labels, and it can be hard to compute. On the other hand, the mean field algorithm for independent posterior component means and labels is easy to iterate.

We will compare risk quantities of the MAP, the posterior mixture, and the mean field mixture. Unlike the single Gaussian location model, we are not able to derive closed-form estimators for the Gaussian mixture model with two or more components, so the calculations will require more numerical approximation and simulation.

Whereas the Gaussian location example analyzed the risk with the data-generating distribution as the first argument of relative entropy, this time we will reverse the order of arguments, placing the estimated distribution first. The resulting risk decomposition comes from the compensation identity (Theorem 1.1); it decomposes the I-risk into I-bias squared and I-variance, according to terminology introduced in Section 1.

The posterior for the component means has density proportional to

$$\left(\prod_k e^{-\frac{1}{2}\|\mu_k\|^2}\right)\left(\prod_i \sum_k e^{-\frac{1}{2\sigma^2}\|X_i-\mu_k\|^2}\right).$$

The mean field algorithm for this context can be adapted from [Bishop, 2006, Sec 10.2]. It alternates updating the "responsibilities" of the various components for the observations and then updating the component means' distributions. The unnormalized responsibility of component j for observation i is

$$\check{\gamma}_{i,j} := e^{-\frac{1}{2\sigma^2}\mathbb{E}\|X_i-\mu_j\|^2} = e^{-\frac{1}{2\sigma^2}[\|X_i-\mathbb{E}\mu_j\|^2+\mathbb{E}\|\mu_j-\mathbb{E}\mu_j\|^2]}$$

where the expectation is with respect to μ_j using its current distribution.[15] For each observation i, the k responsibilities $\gamma_{i,1}, \ldots, \gamma_{i,k}$ are the normalized version of $\check{\gamma}_{i,1}, \ldots, \check{\gamma}_{i,k}$.[16] Once the responsibilities have been calculated, the component mean posteriors are updated as follows. Define n_j to be the sum of the responsibilities of component j, summing over all observations; it can be thought of as the effective sample size for component j. Define also a weighted average for each component by

$$\bar{x}_j := \frac{1}{n_j}\sum_i \gamma_{i,j} x_i.$$

When each component mean's prior is a standard Gaussian, the update rule sets the new distribution for the jth component mean to Gaussian with location $\frac{n_j}{\sigma^2+n_j}\bar{x}_j$ and precision $(\frac{1}{\sigma^2}+n_j)I_d$.

The mean field posterior's component mean distributions remain independent of each other, and its variational posterior mixture is easy to calculate: it is simply the k-component mixture of the individual means' posterior mixtures [Bishop, 2006, Sec 10.2.3]. Each component contributes a continuous Gaussian mixture over a Gaussian location model, which results in a Gaussian; the resulting Gaussian's mean is equal to the posterior's mean while its covariance is equal to the sum of the posterior's covariance and the model covariance. Thus, the mean field mixture is yet another Gaussian mixture, though its component variances differ from each other and are larger than the model variance.

A concrete example will add to our understanding of the relationship between the MAP, the posterior mixture, and the mean field mixture. We use the two-component GRBM model with $\sigma^2 = 1$. Suppose $n = 10$ and the true data-generating distribution is $\frac{1}{2}N(1.75, 1) + \frac{1}{2}N(-1.75, 1)$. Figure 5 (left column) shows the the posterior densities for four different samples. For each sample, a run of the mean field algorithm gives the density shown in the right column.[17] As in the Gaussian location example, it seems reasonable that the mean field mixture can be fruitfully thought of as an estimator somewhere between the true posterior mixture and the MAP; we will explore this idea further with additional plots.

Figure 6 shows the four repetitions of $n = 10$ data points and plots the true data-generating density along with the MAP, the posterior mixture, and

[15] The initializations for the component means' posterior distributions need to be distinct in order to break the symmetry.

[16] This terminology is used by [Bishop, 2006, Sec 10.2], highlighting a parallel with the EM algorithm for mixtures.

[17] In reality, the mean field algorithm only "selects" one of the two modes, but our figure shows an equivalent density since the model is symmetric across the line $\mu_2 = \mu_1$.

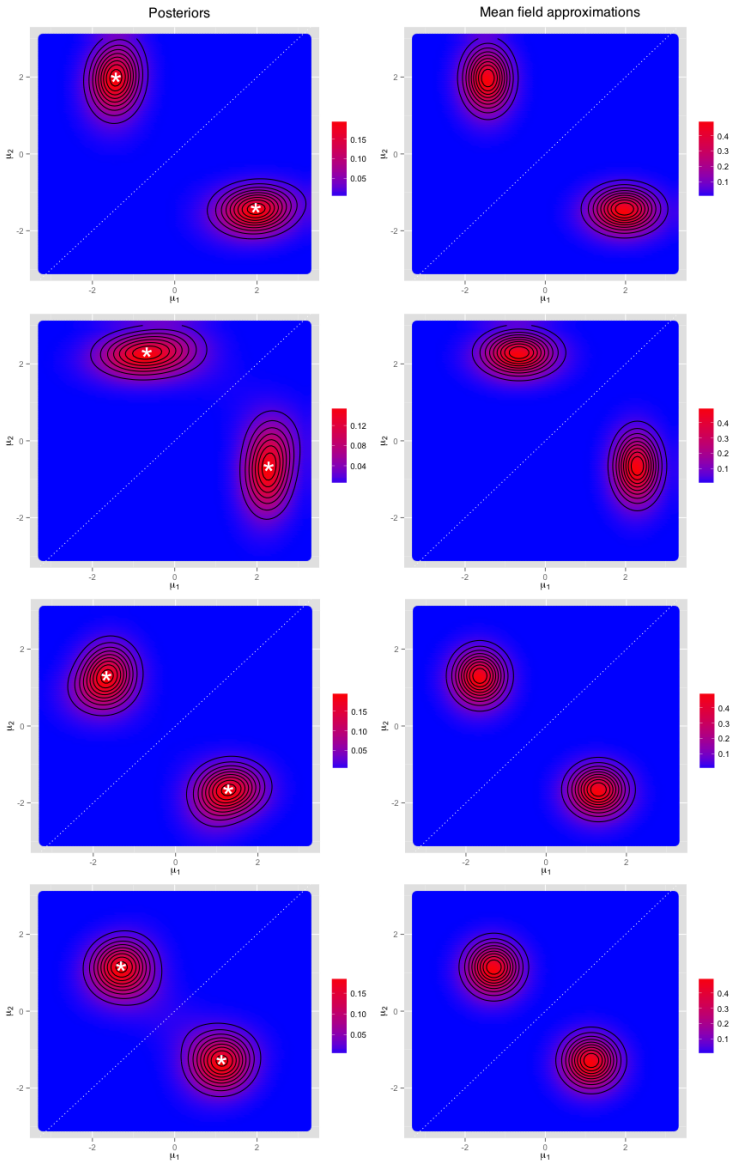

Figure 5: Top left: The standard Gaussian prior on (μ_1, μ_2). Top right: The likelihood of a sample of size $n = 10$ drawn from $\frac{1}{2}N(1.75, 1) + \frac{1}{2}N(-1.75, 1)$. Bottom left: The posterior we get by dividing the product of prior and likelihood by a numerically calculated integral of that product. Bottom right: A (symmetrized) mean field approximation to the posterior according to a run of the mean field algorithm.

the mean field posterior mixture. Each MAP has two distinct peaks, while the posterior mixtures are much more smoothed out, and the mean field mixtures is always between the two.

We also provide Figure 7 to show the centroids of the estimators approximately. To produce these centroids, the MAP, posterior mixture, and mean field mixture were each calculated for $M = 50$ independently generated datasets. Each estimator's centroid is approximated by the uniform mixture of its M realizations. We further refine the approximate centroids by symmetrizing since in this case it is obvious that the centroid should be symmetric about zero. Our figure shows that the mean field mixture's approximate centroid is almost the same as that of the posterior mixture; it is "pulled" a bit toward the MAP's centroid.

Figure 8 shows the three estimates from our first dataset and the approximate centroids. It also reports the estimators' I-variances each of which is approximated by the average of the relative entropies from the $M = 50$ estimates to the centroid.[18]

To see how the MAP, posterior mixture, and mean field posterior mixture compare over a larger set of possible data-generating distributions, we consider probability measures of the form $\frac{1}{2}N(a, s^2) + \frac{1}{2}N(-a, s^2)$. Figure 9 visualizes the three estimators' compensation identity quantities over a range of (a, s).[19]

These examples reinforce the idea that the mean field mixture's behavior is between that of the posterior mixture and the MAP. In Figure 7, the mean field mixture's centroid is almost the same as the posterior mixture's, both of which are better than that of the MAP. That same phenomenon is seen more generally in the top row of Figure 9. Figure 8 indicates that the mean field mixture's I-variance is close to that of the posterior mixture and that both of which are a bit smaller than that of the MAP; again the Figure 9 mirrors this observation in its middle row.

In Bayesian estimation, the posterior mixture is in many ways *statistically preferable* to the MAP.[20] It is also *computationally infeasible* except in certain cases, unlike the MAP. For Bayesian Gaussian mixture modeling, the posterior mixture indeed has no closed form. Our example demonstrates that the mean field approximation can combine the advantages of both estimators: it can gain the statistical benefits of mixing while retaining a simple closed form.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0387-31073-2.

[18] It was necessary to use a relatively small M because the numerical double integrals are time-consuming. With high-quality computing hardware, more precise calculations can be achieved.

[19] For these calculations, we continue to use $\sigma^2 = 1$ for the model; the truth is only in the model when $s = 1$.

[20] See the discussion of the posterior mixture as a Bayes rule in Section ??.

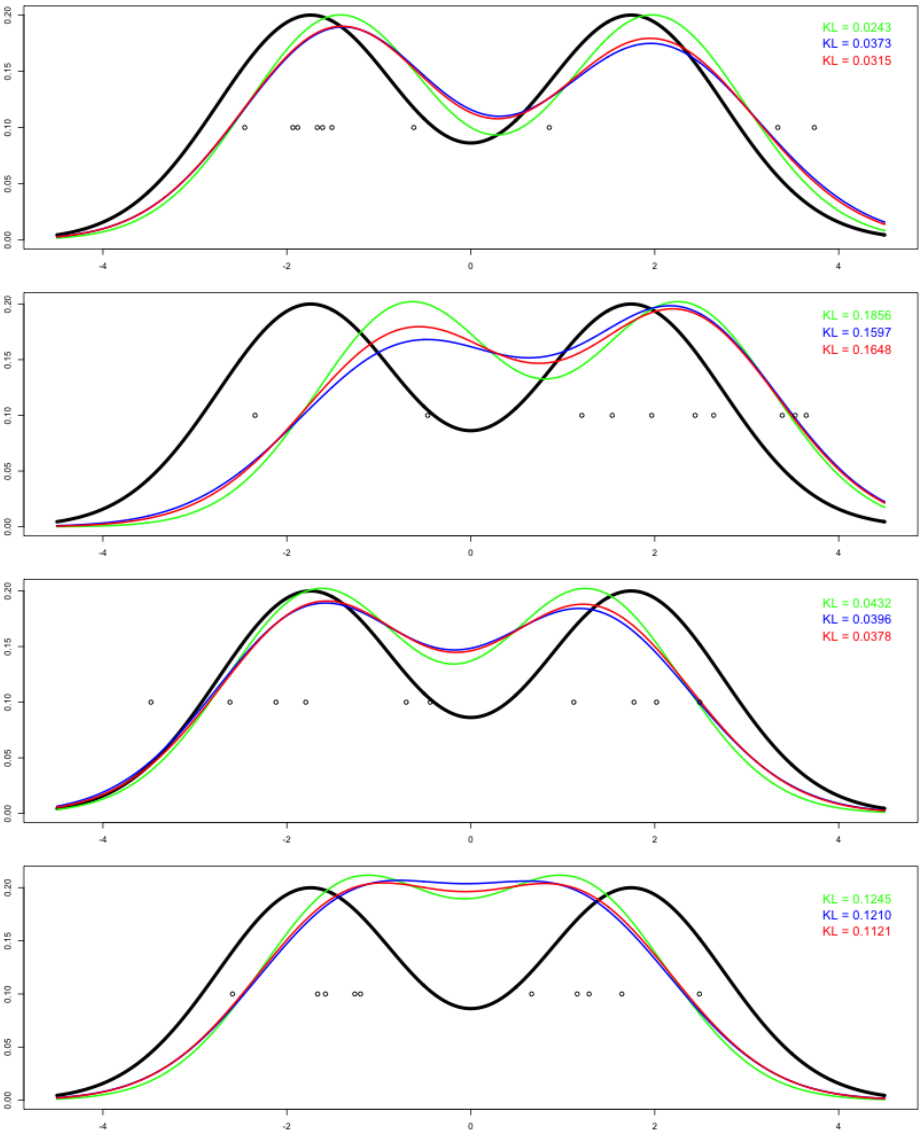

Figure 6: The data-generating distribution $\frac{1}{2}N(1.75, 1) + \frac{1}{2}N(-1.75, 1)$ has density shown in black. After each of four samples of size $n = 10$ (shown by the black points), the MAP (green), posterior mixture (blue), and mean field posterior mixture (red) are plotted and their relative entropies to the data-generating distribution are stated to four decimal places.

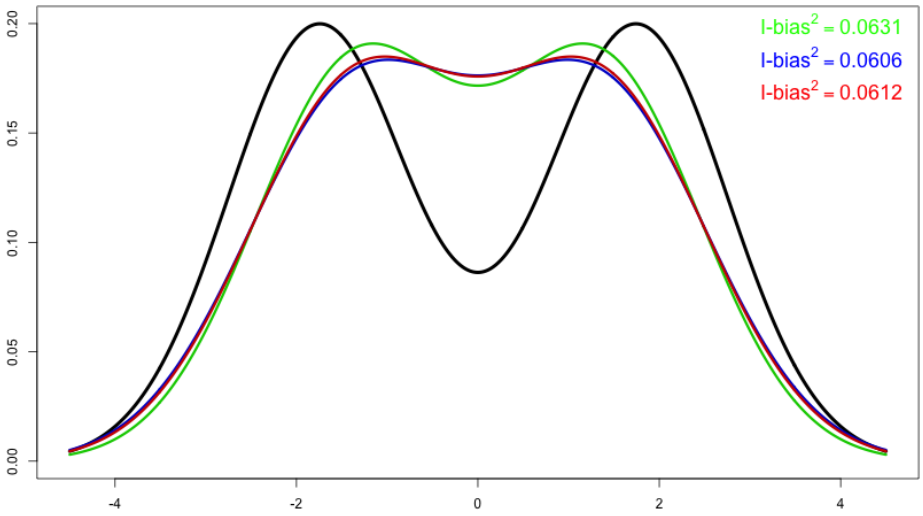

Figure 7: The data-generating distribution $\frac{1}{2}N(1.75, 1) + \frac{1}{2}N(-1.75, 1)$ has density shown in black. For $n = 10$, approximate centroids of MAP (green), posterior mixture (blue), and mean field posterior mixture (red) are shown and their relative entropies to the data-generating distribution (which are the estimators' I-biases squared) are stated to four decimals.

W. D. Brinda. *Adaptive Estimation with Gaussian Radial Basis Mixtures*. PhD thesis, Yale University, 2018.

W. D. Brinda and Jason M. Klusowski. Hölder's identity. Submitted to Bernoulli in 2018.

Zoubin Ghahramani. Factorial learning and the em algorithm. In *Advances in neural information processing systems*, pages 617–624, 1995.

Peter D. Grünwald. *The minimum description length principle*. MIT press, 2007. ISBN 0262072815.

Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

Kevin P. Murphy. Conjugate bayesian analysis of the gaussian distribution. Available at `www.cs.ubc.ca/~murphyk`, 2007.

David Pfau. A generalized bias-variance decomposition for bregman divergences. Available at `davidpfau.com`, 2013.

Matus Telgarsky and Sanjoy Dasgupta. Agglomerative bregman clustering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1011–1018. Omnipress, 2012.

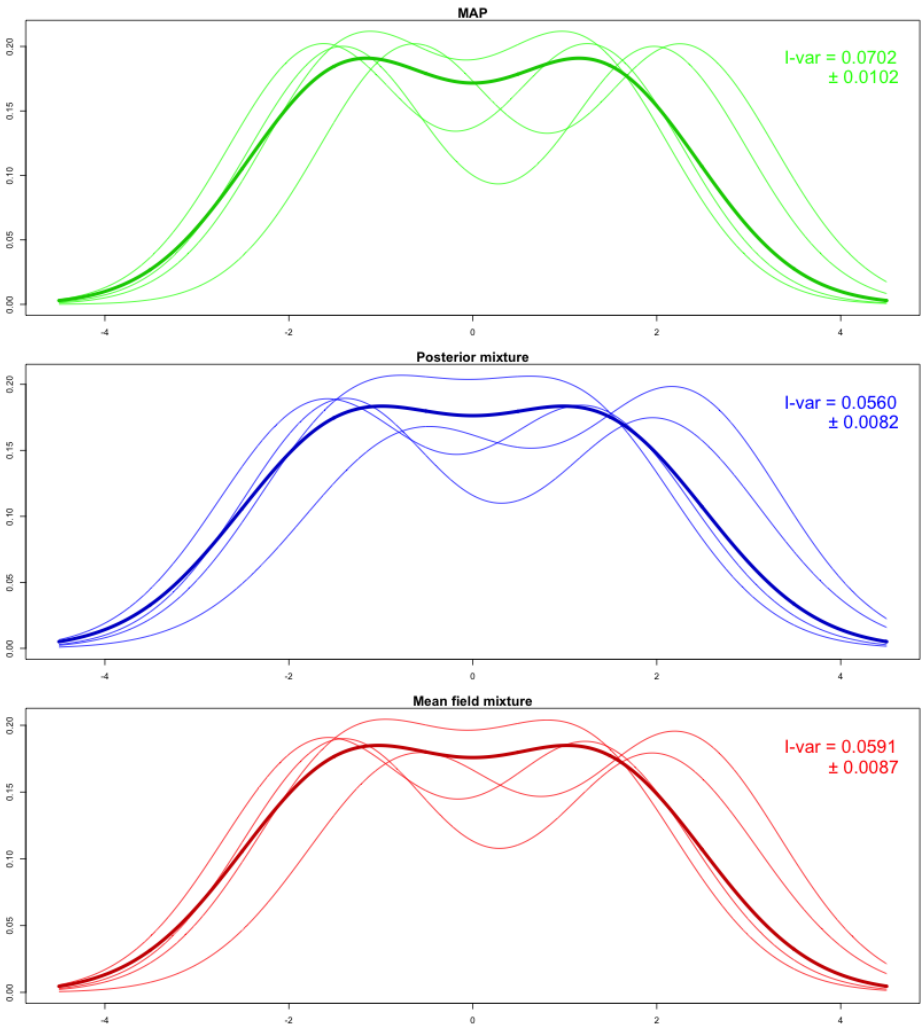

Figure 8: Four repetitions of $n = 10$ draws are taken from the data-generating distribution $\frac{1}{2}N(1.75, 1) + \frac{1}{2}N(-1.75, 1)$. The MAP (top), posterior mixture (middle), and mean field mixture (bottom) are plotted along with their approximate centroids (darker). I-variances are approximated by the average $M = 50$ simulations' estimators from the centroids, and the standard errors arising from the simulated samples are given.

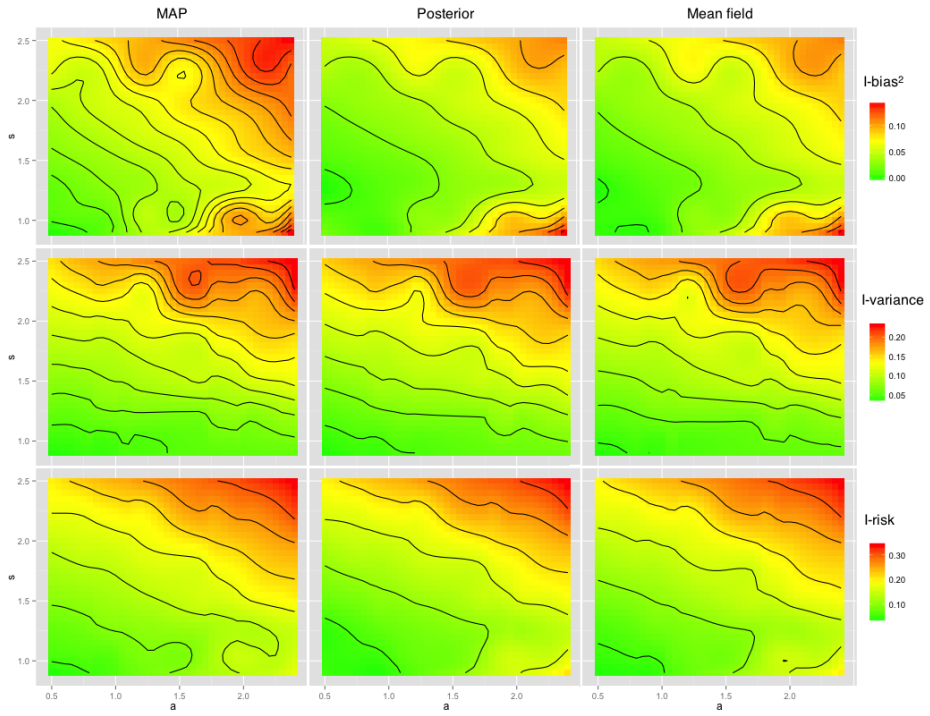

Figure 9: $n = 10$ draws are taken from the data-generating distribution $\frac{1}{2}N(a, s^2) + \frac{1}{2}N(-a, s^2)$. Heat maps show the I-risk quantities of the MAP, posterior mixture, and mean field posterior mixture for a range of (a, s). The values were calculated approximately by simulation, resulting in noisy (and overly correlated) contours even after smoothing.

Flemming Topsøe. Basic concepts, identities and inequalities — the toolkit of information theory. *Entropy*, 3(3):162–190, 2001.