When are Bayesian model probabilities overconfident?

Oscar Oelrich[a], Shutong Ding[b], Måns Magnusson[c], Aki Vehtari[c], and Mattias Villani[a,d]

[a]Department of Statistics, Stockholm University; [b]Department of Statistics, International Labour Organization; [c]Department of Computer Science, Aalto University; [d]Department of Computer and Information Science, Linköping University

This manuscript was compiled on March 10, 2020

Abstract. **Bayesian model comparison is often based on the posterior distribution over the set of compared models. This distribution is often observed to concentrate on a single model even when other measures of model fit or forecasting ability indicate no strong preference. Furthermore, a moderate change in the data sample can easily shift the posterior model probabilities to concentrate on another model. We document overconfidence in two high-profile applications in economics and neuroscience. To shed more light on the sources of overconfidence we derive the sampling variance of the Bayes factor in univariate and multivariate linear regression. The results show that overconfidence is likely to happen when i) the compared models give very different approximations of the data-generating process, ii) the models are very flexible with large degrees of freedom that are not shared between the models, and iii) the models underestimate the true variability in the data.**

Model comparison | Bootstrap | DSGE | Macroeconomic policy | Brain connectivity

Bayesian inference has gained widespread popularity in recent decades, largely propelled by advances in computing power and efficient simulation algorithms like Markov Chain Monte Carlo (1) and Sequential Monte Carlo (2). The Bayesian approach to model comparison is theoretically attractive and the standard in many fields in both the natural and the social sciences. Some examples are (3) in Linguistics, (4) in Economics, and (5) in Neuroscience. Hypothesis testing is a special case of model selection. Problems with classical hypothesis testing and the so called reproducibility crisis has directed attention to Bayesian model selection as an alternative, see for example (6) and (7).

A posterior distribution over a set of models makes it straightforward to select one of the models for further study, or to average inference across the models using Bayesian Model Averaging (BMA, 8). Bayesian model comparison based on posterior model probabilities has many attractive properties: i) it allows the compared models to be non-nested, ii) it is consistent when the data generating process is among the compared models (the $\mathcal{M}$-closed perspective in 9), iii) it will asymptotically concentrate the posterior probability mass on the model closest to the data generating process when all compared models are misspecified ($\mathcal{M}$-open perspective in 9), and iv) it has direct connections to out-of-sample forecasting performance and cross-validation (9–13).

However, Bayesian model probabilities sometimes behave puzzling in practise in that the posterior model distribution often concentrates entirely on one model, giving the impression of overwhelming support for that model. At the same time other forms of model comparison, e.g. predictive performance on a test set, do not show nearly the same degree of discrimination. This overconfidence is part of the folklore among expert Bayesians, but remains largely undocumented in the scientific literature, barring brief passages such as in (14) or (15) who state 'In practice we have observed a tendency of BMA to be over confident in weighting models—assigning weights that are too close to zero or one'. Moreover, theoretical work on overconfidence in Bayesian model probabilities is scarce. Two recent exceptions are (16) and (17). The asymptotic behavior of Bayesian model comparison when the compared models are equally misspecified is explored in (16), showing random walk like behavior of the log Bayes factor in large samples. The overconfidence of Bayesian posteriors and Bayesian model probabilities is also highlighted by (17) who use bagging of posteriors to make Bayesian inference more robust.

Our paper sheds light on the sources of overconfidence of Bayesian model probabilities by deriving the sampling variance of the Bayes factor in linear regression. We show that overconfidence is likely to be happen when i) the compared models give very different approximations of the data-generating process, ii) the models are very flexible, i.e. have large degrees of freedom, and that complexity is not shared between the models, and iii) the models are unable to replicate the variability in the data generating process. We also extend the results to multivariate regression.

The next section motivates our study by showing disturbingly clear signs of overconfidence in a high-profile applications in macroeconomics (4) and neuroscience (18). The rest of the paper studies the sources of overconfidence mathematically through the between-sample variance of the Bayes factor for Bayesian linear regression models. Proofs of the results are given in the Supplementary material.

Significance Statement

Bayesian model inference holds a central place in a wide range of scientific fields. However, Bayesian model probabilities often behave paradoxically in that they tend to concentrate almost entirely on a single model even when other measures of model fit do not indicate large differences between models. We document this overconfidence in high-profile applications in macroeconomics and neuroscience, and characterize the sources of overconfidence in a regression setting.

M.V. proposed the initial idea for the paper and designed research. M.V., O.O. and S.D. proved the mathematical results. M.V., M.M., O.O. and S.D. wrote the computer code, analyzed data and conducted the experiments. M.V., O.O. and S.D. wrote the initial draft of the paper. All authors edited the paper and contributed in discussions.

The authors declare no conflict of interest.

[1]M.V., O.O. and S.D. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: oscar.oelrich@stat.su.se

Background and motivation

Bayesian model probabilities. Consider comparing a set of K models, $\mathcal{M} = \{M_1, ..., M_K\}$, for the observed data $y = (y_1, ..., y_n)^T$, each depending on a vector of model parameters θ_k. A common way of doing Bayesian model comparison is to use the posterior distribution over the model set $\mathcal{M}$,

$$p(M_k|y) \propto p(y|M_k)p(M_k),$$

where $p(M_k)$ is the prior probability of model M_k, $p(y|M_k)$ is the *marginal likelihood*

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k,$$

and $p(\theta_k|M_k)$ is the prior distribution for θ_k. The Bayes factor for comparing model M_k to model M_l is

$$\mathrm{B}_{k,l} = \frac{p(y|M_k)}{p(y|M_l)}.$$

Micro-based general equilibrium models in macroeconomics. To illustrate the effect of overconfidence in real-world applications, we first consider a class of Dynamic Stochastic General Equilibrium (19) models widely used in economics. DSGE models are the main models used for policy analysis and prediction at essentially every major monetary and fiscal institution in the world. Bayesian model probabilities is the standard tool for model comparison and selection among DSGE models (19). The seven-variable Smets-Wouters (4) model is the de facto starting point for most DSGE models used in practical work. The Smets-Wouters model is a probability model for seven macroeconomic time series using a complex micro-funded model based on optimizing representative agents in the economy with rational expectations. The model dynamics are driven by seven underlying latent time series shocks, such as shocks in technology and preferences. In (4), the base version of the model is compared with eight variants that restricts certain model parameters to known values. The posterior model probabilities based on the marginal likelihoods from (4) is given in Table 1, showing conclusive evidence in favor of model M_3.

Table 1. Posterior model probabilities in the DSGE example.

Base	M1	M2	M3	M4	M5	M6	M7	M8
0.01	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00

To investigate if overconfidence is a concern we approximate the sampling distribution of the posterior model probabilities using the circular block boostrap for time series (20). Figure 1 shows results from 1000 bootstrap replicates. The vertical bars correspond to models and the horizontal stripes to bootstrap replicates. The bootstrap replicates have been sorted with respect to the probabilities of the baseline model. The colors represent the posterior model probabilities. For example, a row where one of the columns has a stripe of dark purple implies that the model in that column has strong support ($\Pr(M_k|\text{Data}) > 0.9$), for the given bootstrap replicate, and all other models weak support ($\Pr(M_k|\text{Data}) \leq 0.1$). Figure 1 shows that the conclusion from Table 1, where model M_3 came out as the sure winner, is far from robust. In a large fraction of

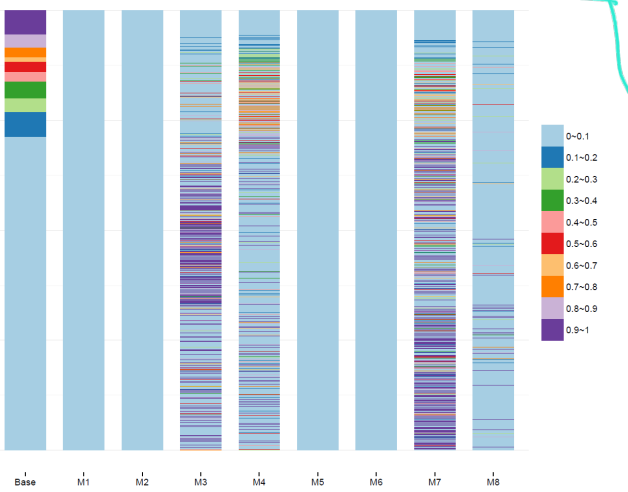

Fig. 1. Posterior model probabilities for the nine DSGE models in (4) for 1000 bootstrap replicates. The bars correspond to models and the horizontal stripes to bootstrap replicates, which have been sorted with respect to the probabilities of the base model. The colors correspond to posterior model probabilities. For example, a row where one of the bars has a stripe of dark purple implies that the model corresponding to that bar has strong support ($p(M_k|y) > 0.9$) for the given bootstrap replicate, and all other models weak support ($p(M_k|y) \leq 0.1$).

Fig. 2. Sampling variability in the DSGE model selection process for three posterior model probability (PMP) thresholds. The figure displays the percentage of bootstrap samples in which one model is strongly preferred, or in which the evidence is inconclusive.

the bootstrap samples, we actually have $\Pr(M_7|\text{Data}) > 0.9$; there is also the same level of support for the base model in a non-negligible fraction of bootstrap replicates.

Figure 2 displays the percentage of bootstrap replicates where there is strong support for one of the models (the posterior model probability, PMP, is larger than 0.9, 0.95 and 0.99, respectively), or where the evidence is inconclusive. If we take the outcome that one of the PMPs are larger than 0.99 as conclusive, then Figure 2 shows that we have conclusive evidence in 35% of the replicates, but spread over 5 different models. We therefore conclude that the model comparison in (4) suffers from overconfidence with misleadingly conclusive support for M_3. The authors of (4) seem in fact unimpressed by the strong support for model M_3 as they silently continue the remainder of the article with the analysis of the base model.

Causal brain interactions in neuroscience. Dynamic Causal Models (DCM) (21) is a popular class of models that use data from functional MRI (fMRI) brain scans to investigate how brain regions interact during an experimental task. Bayesian posterior model probabilities are the recommended method of model comparison for DCMs (22)[Ch.37]. DCMs are used in (18) to analyze how three brain regions that are known to be

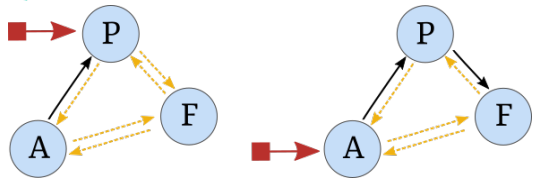

Fig. 3. Two of the compared DCM models for the three brain auditory brain regions: i) left posterior temporal sulcus (P), ii) left anterior superior temporal sulcus (A) and iii) pars orbitalis of the inferior frontal gyru (F). Dashed orange arrows are endogenous task-unrelated connections which are present in all models, while solid black arrows are connections whose strength is modulated by the hearing task. The location of the auditory input is indicated by the red square with an arrow.

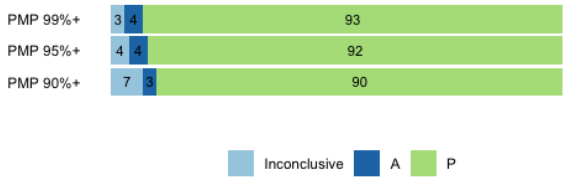

Fig. 4. Sampling variability in the DCM model selection process for three posterior model probability (PMP) thresholds. The figure displays the percentage of bootstrap samples in which one model is strongly preferred, or in which the evidence is inconclusive.

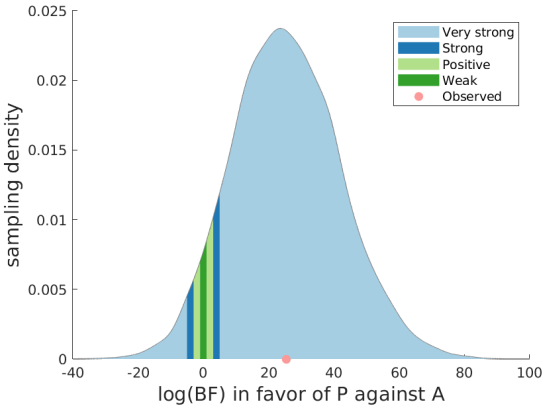

Fig. 5. Bootstrap approximation of the sampling distribution of $\log \mathrm{BF}_{P,A}$ comparing the family of models where the auditory input is localized only in region P against the family of models where it is localized only in region A. The sampling distribution is approximated by 10000 bootstrap resamples of subjects. The Kass-Raftery scale of evidence is displayed by the colored regions. The observed $\log \mathrm{BF}_{P,A}$ is marked out by a pink dot.

associated with speech processing interact when hearing intelligible speech. We reanalyze their data, but exclude subject 5 since it is a duplicate of subject 4. This leaves 25 subjects for the analysis.

Figure 3 displays two of the compared models. A particular question of interest is in which of the three regions the auditory input is localized, regardless of the presence or absence of connectivity patterns. Such posterior probabilities are obtained by summing over all possible connections for a given source location, and is presented in Table 2, which is identical to Table 1 in (18) even with subject 5 removed. According to Table 2 we are supposed to be absolutely certain that input only enters through region P. Figure 4 however shows that in a non-negligible fraction of bootstrap samples we actually obtain conclusive evidence for region A. Figure shows the bootstrapped sampling distribution for the log Bayes factor comparing the hypothesis P against A. The regions of evidence from the well known conservative Kass-Raftery scale (23) are also indicated in the figure. The sampling variance is very large, and the regions of weak or inconclusive support is but a small interval between large masses of very strong support for either of the two hypotheses. Most of the mass for strong evidence happens to be located on P in this example, but it is clear that the Bayes factor can very easily be overconfident.

Table 2. Posterior model probabilities in the DCM example.

A	F	P	AF	PA	PF	PAF
0.00	0.00	1.00	0.00	0.00	0.00	0.00

Overconfidence in Gaussian linear regression models

Univariate response. We explore the sources of overconfidence of Bayesian model comparison by deriving the sampling variance of the Bayes factor for two compared models in a finite sample setting. To get tractable and easily interpretable results we consider Gaussian linear regression models with known error variances σ_1^2 and σ_2^2; Remark 2 discusses the case with unknown variances. The data generating process M_* is given by

$$M_* : \quad y = X_*\beta_* + \varepsilon_* \quad \varepsilon_* \sim \mathrm{N}(0, \sigma_*^2 I_n), \qquad [1]$$

where X_* is a $n \times p_*$ matrix of full rank. We compare the misspecified models

$$\begin{aligned} M_1 : \quad & y = X_1\beta_1 + \varepsilon_1 \quad \varepsilon_1 \sim \mathrm{N}(0, \sigma_1^2 I_n) \\ M_2 : \quad & y = X_2\beta_2 + \varepsilon_2 \quad \varepsilon_2 \sim \mathrm{N}(0, \sigma_2^2 I_n), \end{aligned} \qquad [2]$$

where X_i is known and of full rank p_i, and β_i is a $p_i \times 1$ vector of unknown regression coefficients.

We use Zellner's g-prior for both models

$$\beta_i|\sigma_i^2 \sim \mathrm{N}\left(0, g\sigma_i^2 (X_i^T X_i)^{-1}\right)$$

for some shrinkage constant $g > 0$, but it is straightforward to extend the results to general normal priors. The posterior is of the form

$$\beta_i|\sigma_i^2, y \sim \mathrm{N}\left(\tilde{\beta}_i, \kappa\sigma_i^2 \left(X_i^T X_i\right)^{-1}\right),$$

where

$$\tilde{\beta}_i = \kappa \left(X_i^T X_i\right)^{-1} X_i^T y,$$

and $\kappa = \frac{g}{g+1}$ is the shrinkage factor. The Bayesian posterior predictive mean of model M_i is a linear smoother (24, Ch. 3.10) of the form $\hat{y}_i = H_i y$, where $H_i = \kappa P_i$ is a shrunken version of the least squares projection matrix $P_i = X_i(X_i^T X_i)^{-1} X_i^T$.

The marginal likelihood for model M_i is given by

$$p(y|M_i) = (2\pi\sigma_i^2)^{-\frac{n}{2}} (1-\kappa)^{p_i/2} \exp\left\{-\frac{1}{2\sigma_i^2} y^T (I_n - H_i) y\right\}.$$

Our first result is derived under the assumption that the error variances in the two models are known and equal. This

case gives a particularly simple expression with interesting interpretation. See the Appendix for results when the model variances differ. Let $\|x\|_2 = (x^T x)^{1/2}$ be the Euclidean norm of the vector x and $\|A\|_F = \sqrt{\mathrm{tr}(A^T A)}$ the Frobenius norm of the matrix A.

Theorem 1. *The sampling mean and variance of the log Bayes factor for the two regression models in [2], assuming equal and fixed variance σ^2, with respect to the data-generating process in [1] is*

$$\mathrm{E}(\log \mathrm{B}_{12}(y)) = \frac{\mathrm{KL}_2 - \mathrm{KL}_1}{2-\kappa} + \frac{p_1 - p_2}{2}\left(\log(1-\kappa) + \kappa\frac{\sigma_*^2}{\sigma^2}\right)$$

$$\mathrm{Var}(\log \mathrm{B}_{12}(y)) = \frac{\sigma_*^2}{\sigma^2}\frac{\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2}{\sigma^2} + \left(\frac{\sigma_*^2}{\sqrt{2}\sigma^2}\|H_1 - H_2\|_F\right)^2,$$

where $\hat{\mu}_i = H_i\mu_*$ is the projection of the true mean vector μ_* onto model M_i and KL_i is the Kullback-Leibler divergence of model M_i with estimate $\hat{\mu}_i$ from the true M_*.

Theorem 1 shows that the Bayes factor favors models that are KL-close to the data-generating process, which is in line with the general asymptotic result in for example (25, 26), but also that it penalizes complex models.

More interestingly, Theorem 1 shows that the variance increases with: i) $\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2/\sigma^2$, ii) $\|H_1 - H_2\|_F$, and iii) the variance ratio σ_*^2/σ^2. We discuss each of these parts in turn.

It is straightforward to prove that $\hat{\mu}_i$ minimizes the KL divergence of $p(y|M_i)$ from $p(y|M_*)$ (see the Appendix); hence, $\hat{\mu}_i$ is the best approximation of $p(y|M_*)$ that model M_i is capable of. It is also easy to show (see the Appendix) that $\mathrm{KL}(M_1(\hat{\mu}_1)\,\|\,M_2(\hat{\mu}_2)) = \|\hat{\mu}_1 - \hat{\mu}_2\|_2^2/2\sigma^2$, the KL divergence between the best approximating models $\mathrm{N}(y|\hat{\mu}_1, \sigma^2)$ and $\mathrm{N}(y|\hat{\mu}_2, \sigma^2)$, which happens to be symmetric when $\sigma_1 = \sigma_2$. The first term in Theorem 1 therefore shows that the variance of the Bayes factor tends to be large when the two models approximate M_* in widely different ways. This explains why continuous model expansion, where a model is embedded in a larger family via a continuous parameter, is preferred over a comparison of a discrete set of well separated models (27). A model is always surrounded by other similar models in continuous model expansions. We also note that the recommended strategy in (27) is to compare widely different models that 'stake out the corners in the model space' in order to capture the true model uncertainty. This is a good strategy when one can afford to stake out the corners with a dense set of models, preferably even a continuum of models, but this is rarely the case. A much more common situation is when the model space is staked out using a small set of models. Unfortunately, Theorem 1 shows that the posterior model probabilities are then highly likely to be overconfident.

To interpret the term $\|H_1 - H_2\|_F$ recall that the *degrees of freedom* of a linear smoother, $\hat{y} = Hy$, is given by $\mathrm{tr}(H)$ (28). The next lemma shows that $\|H_1 - H_2\|_F^2$ is the total degrees of freedom of the two models that is not shared between them.

Lemma 1. *$\|H_1 - H_2\|_F^2$ measures the total non-shared degrees of freedom of the models in [2]:*

$$\|H_1 - H_2\|_F^2 = \kappa^2\left(\mathrm{tr}(P_1) + \mathrm{tr}(P_2) - 2\left(s + \Sigma_{i=1}^r \cos^2(\theta_{k+i})\right)\right),$$

where $p_1 = \mathrm{tr}(P_1) \geq \mathrm{tr}(P_2) = p_2 \geq 1, \theta_j \in [0, \pi/2]$ for $j = 1, ..., p_2$ are the principal angles between $S_1 = \mathrm{span}(H_1)$ and $S_2 = \mathrm{span}(H_2)$, $s = dim\,(S_1 \cap S_2)$ is the number of θ_j that are exactly zero and r is the number of θ_j in the open interval $(0, \pi/2)$.

By Lemma 1, $\|H_1 - H_2\|_F^2$ is the total complexity of the two models, $\mathrm{tr}(P_1)+\mathrm{tr}(P_2)$, reduced by the s completely shared dimensions and by the cosine of the principal angles of the r partially shared dimensions, shrunk by the precision of the prior. Hence, $\mathrm{Var}(\log \mathrm{B}_{12}(y))$ is not affected by any complexity that is shared between the models. Comparison of a dense set of nested models, e.g. variable selection in regression, is therefore expected to less prone to overconfidence since there is often a large overlap between compared models. Figure 6 gives an abstract illustration of models, divergences between models and their shared/non-shared complexities.

Finally, both $\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2/\sigma^2$ and $\|H_1 - H_2\|_F$ in Theorem 1 are inflated by the error variance ratio σ_*^2/σ^2 in the expression for $\mathrm{Var}(\log \mathrm{B}_{12}(y))$. Models that are unable to generate enough variation in their data distribution are therefore particularly susceptible to overconfidence.

We make the following additional remarks about Theorem 1.

Remark 1. $\mathrm{Var}(\log \mathrm{B}_{12}(y))$ *does not directly depend on the degree of misspecification of the two models relative to the true M_*; only the divergence between the models matters. However, the degree of misspecification restricts how far apart the models can be. One way to see this is by noting that $\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2$ has the upper bound*

$$\|\hat{\mu}_* - \hat{\mu}_1\|_2^2 + \|\hat{\mu}_* - \hat{\mu}_2\|_2^2 + 2\,\|\hat{\mu}_* - \hat{\mu}_1\|_2^2\,\|\hat{\mu}_* - \hat{\mu}_2\|_2^2,$$

which under equal degree of misspecification simplifies to

$$\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2 \leq 4\,\|\hat{\mu}_* - \hat{\mu}_i\|_2^2.$$

Hence, in problems where all models are very misspecified there is greater scope for the models to approximate $p(y|M_*)$ in markedly different ways, and therefore greater risk of overconfidence.

Remark 2. *Since the known variance σ^2 enters the mean and variance in Theorem 1 as a multiplicative factor $1/\sigma^2$, it is straightforward to generalize Theorem 1 to the case with a common and unknown variance by applying the law of total variance. The end result is that the factor $1/\sigma^2$ is replaced by its prior expectation, but the two main terms $\|\hat{\mu}_1 - \hat{\mu}_2\|_2^2/\sigma^2$ and $\|H_1 - H_2\|_F$ remain. The case with unknown and different variances seems to be intractable, but numerical experiments indicate that the same three factors are driving $\mathrm{Var}(\log \mathrm{B}_{12}(y))$.*

Remark 3. *The technique behind Theorem 1 can also be used for regression models with heteroscedastic variance as the log of the marginal likelihood remains quadratic in y with $I_n - H_i$ replaced by $\Sigma_\varepsilon^{-1/2}(I_n - H_i)\Sigma_\varepsilon^{-1/2}$, where Σ_ε is the $n \times n$ covariance matrix of errors. This framework includes the popular Gaussian process regression models in machine learning where*

$$y_i = x_i^T\beta + f(x),$$

and the function $f(x)$ follows a Gaussian process with a given covariance kernel. The log marginal likelihood is given in Equation 2.43 of (29).

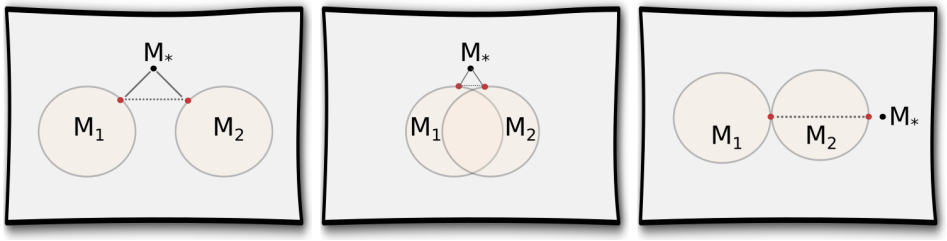

Fig. 6. Illustrating divergence and complexity of models. Each figure displays the data generating process (M_*) as a black point. The two models M_1 and M_2 are depicted as circles where the area of each circle indicates the expressiveness, or complexity, of the model, i.e. $\mathrm{tr}(H_i)$. The red points represent $\hat{\mu}_i$, the best approximation to M_* within M_i. Finally, the dotted line illustrates the Kullback-Leibler divergence between the models, i.e. $||\hat{\mu}_1 - \hat{\mu}_2||$. The left graph shows models with no shared complexity, the middle graph shows models with substantial shared complexity, and the graph to the right shows models without shared complexity where one of the models is closer to M_*.

Multivariate response. The previous section shows that overconfidence is a bigger concern when the compared models are approximating the data-generating process in very different ways. The DSGE models from (4) are seemingly similar, however; they only differ from the baseline model by setting a single parameter to a specific value. Note however that the results in the previous section are derived for the case with a univariate response, while the DSGE models are multivariate model with seven time series responses. This subsection extends the previous results to multivariate regression and highlights some special properties for this more general case.

The data-generating process M_* is given by the multivariate regression

$$M_*: \qquad Y = X_* B_* + E_* \quad E_* \sim \mathrm{N}(0, I_n, \Sigma_*), \qquad [3]$$

and we consider comparing the following models

$$\begin{aligned} M_1: &\qquad Y = X_1 B_1 + E_1 \quad E_1 \sim \mathrm{N}(0, I_n, \Sigma_1) \\ M_2: &\qquad Y = X_2 B_2 + E_2 \quad E_2 \sim \mathrm{N}(0, I_n, \Sigma_2), \end{aligned} \qquad [4]$$

where $\mathrm{N}(\mu, A_1, A_2)$ denotes the matrix variate distribution, Y is $n \times q$, X_i is known and of full rank p_i, and B_i is a $p_i \times q$ vector of unknown parameters. The error term $E_i = (\varepsilon_{1i}, \varepsilon_{2i}, ..., \varepsilon_{ni})'$ is an $n \times q$ matrix, following a matrix normal distribution, with rows that are iid $\mathrm{N}(0, \Sigma_i)$. We let $\beta_i = \mathrm{vec}(B_i)$, and use Zellner's g-prior for the regression coefficients.

Theorem 2. *The sampling mean and variance of the log Bayes factor for the two multivariate regression models in* [4] *with respect to the data-generating process in* [3] *is*

$$\begin{aligned} \mathrm{E}(\log \mathrm{B}_{12}(y)) &= \frac{\mathrm{KL}_2 - \mathrm{KL}_1}{2 - \kappa} \\ &\quad + \frac{p_1 - p_2}{2}\left(\log(1-\kappa) + \kappa \mathrm{tr}\left(\Sigma^{-1}\Sigma_*\right)\right) \\ \mathrm{Var}(\log \mathrm{B}_{12}(y)) &= \frac{1}{2}\mathrm{tr}\left(\Omega^2\right)\|H_2 - H_1\|_F^2 \\ &\quad + \left\|(\hat{\mu}_2 - \hat{\mu}_1)\Sigma^{-1/2}\Omega^{1/2}\right\|_F^2, \end{aligned}$$

where $\Omega \equiv \Sigma^{-1/2}\Sigma_*\Sigma^{-1/2}$ *is a multivariate generalization of the variance ratio* σ_*^2/σ^2 *and* $\hat{\mu}_i = H_i \mu_*$ *is the projection of the true mean vector* μ_* *onto model* M_i.

The interpretation remains largely the same as in the univariate case, with the added insight that not all differences in the models are equally important due to the appearance of the generalized variance ratio Ω in $\left\|(\hat{\mu}_2 - \hat{\mu}_1)\Sigma^{-1/2}\Omega^{1/2}\right\|_F^2$. To show this more precisely, we perform the two spectral decompositions:

$$\begin{aligned} \Sigma &= U\Lambda U^T = \left(U\Lambda^{1/2}\right)\left(U\Lambda^{1/2}\right)^T \\ \Sigma_* &= U_*\Lambda_* U_*^T = \left(U_*\Lambda_*^{1/2}\right)\left(U_*\Lambda_*^{1/2}\right)^T, \end{aligned}$$

where $U = (u_1, ..., u_p)$ and $U_* = (u_{*1}, ..., u_{*p})$ are matrices of eigenvectors, and $\Lambda = \mathrm{Diag}(\lambda_1, ..., \lambda_p)$ and $\Lambda_* = \mathrm{Diag}(\lambda_{*1}, ..., \lambda_{*p})$ are diagonal matrices of eigenvalues. Now, $\Sigma^{-1/2} = U\Lambda^{-1/2}$ and $\Omega = \Lambda^{-1/2}U^T U_*\Lambda_* U_*^T U\Lambda^{-1/2}$, therefore

$$(\hat{\mu}_2 - \hat{\mu}_1)\Sigma^{-1/2}\Omega^{1/2} = (\hat{\mu}_2 - \hat{\mu}_1)U\Lambda^{-1/2}\left(\Lambda^{-1/2}U^T U_*\Lambda_*^{1/2}\right).$$

The $n \times p$ matrix $(\hat{\mu}_1 - \hat{\mu}_2)\,U\Lambda^{-1/2}$ contains the differences in predictions in the directions of the principal components of Σ, rescaled to unit variance. The $p \times p$ matrix $\Lambda^{-1/2}U^T U_*\Lambda_*^{1/2}$ has element $\sqrt{\lambda_{*j}/\lambda_i}\cos(\phi_{u_i, u_{*j}})$ in its ith row, jth column, where $\cos(\phi_{u_i, u_{*j}}) = u_i^T u_{*j}$ measures the degree of alignment of pairs of eigenvectors from Σ and Σ_*. Hence, $\mathrm{var}\log B_{12}(Y)$ will be large when the models make very different prediction on linear combinations of response variables where the eigenvectors of Σ and Σ_* align, and the variance ratio λ_{*j}/λ_i is large. This agrees with the analysis of forecasting performance in (30) for an open-economy extension of the Smets-Wouters model. In addition, (30) show that a multivariate measure of out-of-sample forecasting performance is almost entirely driven by forecasting errors in employment, one of the least important variables from a central bank perspective.

Discussion and Conclusion

We have demonstrated that Bayesian posterior model probabilities can be overconfident in the sense of spuriously picking out one of the compared models as the only probable model in a set of compared models, while at the same time being equally certain about another model in a slightly different dataset. We have analyzed the sources of this overconfidence by deriving the sample variance of the log Bayes factor for univariate and multivariate regression.

The main message is that overconfidence is likely to be a problem when the compared models give very different approximations of the data-generating process and when the compared models are flexible in a way that is not shared among the models. The same is true for the multivariate setting, with the added nuance that overconfidence will be

largest when the models are different with respect to specific linear combinations of the response variables.

Our results motivate several interesting avenues for future research. First, the linear regression setup was chosen since it provides a clear view of what drives overconfidence. It would be illuminating to derive similar measures for more general models to see if the same effects appear there. Second, it would of interest to repeat our analysis for other Bayesian model inference methods, for example prediction pools or stacking (14, 31). Stacking is particularly interesting since it is known to not necessarily concentrate on a single model when the sample grows large. Third, we have used the bootstrap to approximate the sampling distribution of Bayes factors. This can be very time-consuming when models are analyzed by Markov Chain Monte Carlo (MCMC), since we would have to run the MCMC for each bootstrap sample. It would therefore be of practical importance to explore the efficiency of methods where the marginal likelihood for each bootstrap sample is obtained by reweighting the posterior draws from a single MCMC run on the original dataset (32).

1. S Brooks, A Gelman, G Jones, XL Meng, *Handbook of Markov Chain Monte Carlo*. (CRC press), (2011).
2. A Doucet, S Godsill, C Andrieu, On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. computing* **10**, 197–208 (2000).
3. TL Griffiths, M Steyvers, Finding scientific topics. *Proc. Natl. academy Sci.* **101**, 5228–5235 (2004).
4. F Smets, R Wouters, Shocks and frictions in US business cycles: A Bayesian DSGE approach. *Am. Econ. Rev.* (2007).
5. KE Stephan, WD Penny, J Daunizeau, RJ Moran, KJ Friston, Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).
6. AE Raftery, Bayesian model selection in social research. *Sociol. methodology* **25**, 111–164 (1995).
7. VE Johnson, Revised standards for statistical evidence. *Proc. Natl. Acad. Sci.* **110**, 19313–19317 (2013).
8. JA Hoeting, D Madigan, AE Raftery, CT Volinsky, Bayesian Model Averaging: A Tutorial. *Stat. Sci.* **14**, 382–417 (1999).
9. JM Bernardo, AF Smith, *Bayesian Theory*. p. 604 (1994).
10. AE Gelfand, DK Dey, Bayesian model choice: asymptotics and exact calculations. *J. Royal Stat. Soc. Ser. B* (1994).
11. J Geweke, Using simulation methods for Bayesian econometric models: inference, development, and communication. *Econom. reviews* **18**, 1–73 (1999).
12. A Vehtari, J Ojanen, A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6**, 142–228 (2012).
13. E Fong, C Holmes, On the marginal likelihood and cross-validation. *arXiv preprint arXiv:1905.08737* (2019).
14. Y Yao, A Vehtari, D Simpson, A Gelman, Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis* **13**, 917–1003 (2018).
15. M Li, DB Dunson, Comparing and weighting imperfect models using d-probabilities. *J. Am. Stat. Assoc.*, 1–26 (2019).
16. Z Yang, T Zhu, Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci.* **115**, 1854–1859 (2018).
17. JH Huggins, JW Miller, Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104* (2019).
18. AP Leff, et al., The cortical dynamics of intelligible speech. *J. Neurosci.* **28**, 13209–13215 (2008).
19. EP Herbst, F Schorfheide, *Bayesian estimation of DSGE models*. (Princeton University Press), (2015).
20. DN Politis, JP Romano, The stationary bootstrap. *J. Am. Stat. association* **89**, 1303–1313 (1994).
21. KJ Friston, L Harrison, W Penny, Dynamic causal modelling. *Neuroimage* **19**, 1273–1302 (2003).
22. J Ashburner, et al., Spm12 manual. *Wellcome Trust. Centre for Neuroimaging, London, UK*, 2464 (2014).
23. RE Kass, AE Raftery, Bayes factors. *J. Am. Stat. Assoc.* (1995).
24. D Ruppert, MP Wand, RJ Carroll, *Semiparametric regression*. (Cambridge university press) No. 12, (2003).
25. RH Berk, Limiting Behavior of Posterior Distributions when the Model is Incorrect. *The Annals Math. Stat.* **37**, 51–58 (1966).
26. J Fernández-Villaverde, JF Rubio-Ramírez, Comparing dynamic equilibrium models to data: a bayesian approach. *J. Econom.* **123**, 153–187 (2004).
27. D Draper, Assessment and propagation of model uncertainty. *J. Royal Stat. Soc. Ser. B (Methodological)* **57**, 45–70 (1995).
28. T Hastie, R Tibshirani, J Friedman, *The Elements of Statistical Learning*. (2009).
29. CE Rasmussen, Gaussian processes in machine learning in *Summer School on Machine Learning*. (Springer), pp. 63–71 (2003).
30. M Adolfson, J Lindé, M Villani, Forecasting performance of an open economy DSGE model. *Econom. Rev.* **26**, 289–328 (2007).
31. J Geweke, G Amisano, Prediction with misspecified models. *Am. Econ. Rev.* **102**, 482–486 (2012).
32. J Geweke, Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication. *Econom. Revies* **18**, 1–126 (1999).
33. Y Bao, A Ullah, Expectation of quadratic forms in normal and nonnormal variables with applications. *J. Stat. Plan. Inference* **140**, 1193–1205 (2010).
34. TM Cover, JA Thomas, *Elements of information theory*. (John Wiley & Sons), (2012).
35. A Galanti, Subspaces, angles and pairs of orthogonal projections. *Linear Multilinear Algebr.* **56**, 227–260 (2008).

1. Appendix A: Proofs

Preliminaries. The following lemma about Gaussian quadratic forms will be useful.

Lemma 2. *Let A_1 and A_2 be $n \times n$ symmetric, nonstochastic matrices, and let y be an $n \times 1$ vector following a normal distribution, $y \sim \mathrm{N}(\mu, \Sigma)$. Then*

i) $\mathrm{E}(y^T A_1 y) = \mu^T A_1 \mu + \mathrm{tr}(A_1\Sigma)$;

ii) $\mathrm{Var}(y^T A_1 y) = 2\mathrm{tr}((A_1\Sigma)^2) + 4\mu^T A_1 \Sigma A_1 \mu$;

iii) $\mathrm{Cov}(y^T A_1 y, y^T A_2 y) = 2\mathrm{tr}(A_1\Sigma A_2 \Sigma) + 4\mu^T A_1 \Sigma A_2 \mu$.

Proof. Part i) is a standard result and ii) follows from iii). To derive the covariance in iii) note that

$$\mathrm{Cov}(y^T A_1 y, y^T A_2 y) = \mathrm{E}\big((y^T A_1 y)(y^T A_2 y)\big) - \mathrm{E}(y^T A_1 y)\mathrm{E}(y^T A_2 y)$$

and

$$\mathrm{E}\big((y^T A_1 y)(y^T A_2 y)\big) = \mathrm{E}\big((z^T B_1 z)(z^T B_2 z)\big),$$

where $z = \Sigma^{-1/2} y \sim \mathrm{N}(\mu_z, I)$, $\mu_z = \Sigma^{-1/2}\mu$ and $B_j = \Sigma^{1/2} A_j \Sigma^{1/2}$. From Theorem 1 in (33) we have

$$\mathrm{E}\big((z^T B_1 z)(z^T B_2 z)\big) = E(z^T B_1 z)E(z^T B_2 z) + 2\mathrm{tr}(B_1 B_2) + 4\mu_z^T B_1 B_2 \mu_z,$$

from which Lemma 2 iii) follows. □

Proof of Theorem 1. The log Bayes factor is equal to

$$\frac{n}{2}\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{p_1 - p_2}{2}\log(1-\kappa) + Q_2(y) - Q_1(y),$$

where $Q_i(y) = \frac{1}{2\sigma_i^2} y^T (I_n - H_i) y$ and $H_i = \kappa P_i$ with $P_i = X_i \left(X_i^T X_i\right)^{-1} X_i^T$ being a symmetric and idempotent projection matrix. Using Part i of Lemma 2, we have

$$\begin{aligned}\mathrm{E}(\log \mathrm{B}_{12}(y)) =& \frac{n}{2}\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{p_1 - p_2}{2}\log(1-\kappa) + \frac{\sigma_*^2(n-\kappa p_2)}{2\sigma_2^2} - \frac{\sigma_*^2(n-\kappa p_1)}{2\sigma_1^2} \\ &+ \frac{1}{2\sigma_2^2}\mu_*^T(I_n - H_2)\mu_* - \frac{1}{2\sigma_1^2}\mu_*^T(I_n - H_1)\mu_*\end{aligned}$$

since $\mathrm{tr}(\sigma_*^2(I_n - H_i)) = \sigma_*^2(n - \kappa p_i)$. The term $\mu_*^T(I_n - H_i)\mu_*$ can be shown to be a linearly increasing function of the Kullback-Leibler divergence of M_i with the ideal estimate $\hat{\mu}_i = H_i \mu_*$ from M_*:

$$\mathrm{KL}(M_*||M_i) \equiv \int \log\left(\frac{p(y|\mu_*)}{p(y|\hat{\mu}_i)}\right) p(y|\mu_*)dy = -\frac{n}{2}\log\left(\frac{\sigma_*^2}{\sigma_i^2}\right) + \frac{n}{2}\frac{\sigma_*^2}{\sigma_i^2} + \frac{1}{2}\frac{\|\mu_* - \hat{\mu}_i\|_2^2}{\sigma_i^2} - \frac{n}{2},$$

using the Kullback-Leibler divergence between two multivariate normal densities (34). To show the exact connection between $\mathrm{E}(\log \mathrm{B}_{12}(y))$ and $\mathrm{KL}(M_*||M_2) - \mathrm{KL}(M_*||M_1)$ we will here consider the algebraically less involved special case $\sigma_1 = \sigma_2$. For this case, the difference in KL divergences simplifies to

$$\mathrm{KL}(M_*||M_2) - \mathrm{KL}(M_*||M_1) = \frac{1}{2}\frac{\|\mu_* - \hat{\mu}_1\|_2^2}{\sigma^2} - \frac{1}{2}\frac{\|\mu_* - \hat{\mu}_2\|_2^2}{\sigma^2}.$$

Note that

$$\|\mu_* - \hat{\mu}_i\|_2^2 = \mu_*^T\mu_* + \kappa^2\mu_*^T P_i \mu_*^T - 2\kappa\mu_*^T P_i \mu_*^T = \mu_*^T\mu_* - \kappa(2-\kappa)\mu_*^T P_i \mu_*^T$$

so

$$\mu_*^T P_i \mu_*^T = \frac{\mu_*^T\mu_* - \|\mu_* - \hat{\mu}_i\|_2^2}{\kappa(2-\kappa)}.$$

When $\sigma_1 = \sigma_2$ we therefore have that

$$\begin{aligned}\mathrm{E}(\log \mathrm{B}_{12}(y)) &= \frac{p_1 - p_2}{2}\left(\log(1-\kappa) + \kappa\frac{\sigma_*^2}{\sigma^2}\right) + \frac{\kappa}{2\sigma^2}(\mu_*^T P_1 \mu_* - \mu_*^T P_2 \mu_*) \\ &= \frac{p_1 - p_2}{2}\left(\log(1-\kappa) + \kappa\frac{\sigma_*^2}{\sigma^2}\right) + \frac{1}{2-\kappa}\left(\mathrm{KL}(M_*||M_2) - \mathrm{KL}(M_*||M_1)\right).\end{aligned}$$

By Part ii and iii of Lemma 2 the variance is

$$\begin{aligned}
\mathrm{Var}(\log \mathrm{B}_{12}(y)) &= \mathrm{Var}(Q_2(y)) + \mathrm{Var}(Q_1(y)) - 2\mathrm{Cov}(Q_2(y), Q_1(y)) \\
&= \frac{\sigma_*^4}{2\sigma_2^4}\mathrm{tr}((I_n - H_2)^2) + \frac{\sigma_*^2}{\sigma_2^4}\mu_*^T(I_n - H_2)^2\mu_* + \frac{\sigma_*^4}{2\sigma_1^4}\mathrm{tr}((I_n - H_1)^2) + \frac{\sigma_*^2}{\sigma_1^4}\mu_*^T(I_n - H_1)^2\mu_* \\
&\quad - \frac{\sigma_*^4}{\sigma_1^2\sigma_2^2}\mathrm{tr}((I_n - H_2)(I_n - H_1)) - \frac{2\sigma_*^2}{\sigma_1^2\sigma_2^2}\mu_*^T(I_n - H_2)(I_n - H_1)\mu_* \\
&= \frac{\sigma_*^4}{2}\mathrm{tr}\left(\left(\frac{(I_n - H_2)}{\sigma_2^2} - \frac{(I_n - H_1)}{\sigma_1^2}\right)^2\right) + \sigma_*^2\left(\mu_*^T\left(\frac{(I_n - H_2)}{\sigma_2^2} - \frac{(I_n - H_1)}{\sigma_1^2}\right)^2\mu_*\right) \\
&= \frac{\sigma_*^4}{2}\mathrm{tr}\left(\left(\frac{(I_n - H_2)}{\sigma_2^2} - \frac{(I_n - H_1)}{\sigma_1^2}\right)^2\right) \\
&\quad + \sigma_*^2\left(\frac{1}{\sigma_2^2}(\mu_* - \hat{\mu}_2) - \frac{1}{\sigma_1^2}(\mu_* - \hat{\mu}_1)\right)^T\left(\frac{1}{\sigma_2^2}(\mu_* - \hat{\mu}_2) - \frac{1}{\sigma_1^2}(\mu_* - \hat{\mu}_1)\right).
\end{aligned}$$

Assuming that the error variances of the two misspecified models are equal, the expression simplifies to

$$\begin{aligned}
\mathrm{Var}(\log \mathrm{B}_{12}(y)) &= \frac{\sigma_*^4}{2\sigma^4}\mathrm{tr}\left((H_1 - H_2)^2\right) + \frac{\sigma_*^2}{\sigma^4}\left(\mu_*^T(H_1 - H_2)^2\mu_*\right) \\
&= \left(\frac{\sigma_*^2}{\sqrt{2}\sigma^2}\|H_1 - H_2\|_F\right)^2 + \left(\frac{\sigma_*}{\sigma}\frac{\|\hat{\mu}_1 - \hat{\mu}_2\|_2}{\sigma}\right)^2.
\end{aligned}$$

A. Proof of Theorem 2. We derive the result for the case $\Sigma_1 = \Sigma_2$. The proof when $\Sigma_1 \neq \Sigma_2$ is similar, but more tedious algebraically. The log Bayes factor is

$$\begin{aligned}
\log \mathrm{B}_{12}(y) &= \frac{p_1 - p_2}{2}\log(1-\kappa) - \frac{1}{2}\mathrm{tr}\left(Y^T\left(I_n - H_1\right)Y\Sigma^{-1}\right) + \frac{1}{2}\mathrm{tr}\left(Y^T\left(I_n - H_2\right)Y\Sigma^{-1}\right) \\
&= \frac{p_1 - p_2}{2}\log(1-\kappa) - \frac{1}{2}\mathrm{tr}\left(Y^T\left(H_2 - H_1\right)Y\Sigma^{-1}\right) \\
&= \frac{p_1 - p_2}{2}\log(1-\kappa) - \frac{1}{2}\mathrm{vec}(Y)^T\left(\Sigma^{-1}\otimes\left(H_2 - H_1\right)\right)\mathrm{vec}(Y),
\end{aligned}$$

where $\mathrm{vec}(Y) \sim \mathrm{N}\left(\mathrm{vec}(\mu_*), \Sigma_* \otimes I_n\right)$, $\mu_* = X_* B_*$, and $H_i = \kappa P_i$ with $P_i = X_i\left(X_i^T X_i\right)^{-1} X_i^T$. Using Lemma 2, we obtain the mean as

$$\begin{aligned}
\mathrm{E}(\log \mathrm{B}_{12}(Y)) &= \frac{p_1 - p_2}{2}\log(1-\kappa) - \frac{1}{2}(\mathrm{vec}(\mu_*))^T\left(\Sigma^{-1}\otimes\left(H_2 - H_1\right)\right)\mathrm{vec}(\mu_*) - \frac{1}{2}\mathrm{tr}\left(\left(\Sigma^{-1}\otimes\left(H_2 - H_1\right)\right)\Sigma_*\otimes I_n\right) \\
&= \frac{p_1 - p_2}{2}\log(1-\kappa) - \frac{1}{2}\mathrm{tr}\left(\mu_*^T\left(H_2 - H_1\right)\mu_*\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\otimes\left(H_2 - H_1\right)\right) \\
&= \frac{p_1 - p_2}{2}\log(1-\kappa) + \frac{1}{2}\mathrm{tr}\left(\mu_*^T H_1\mu_*\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(\mu_*^T H_2\mu_*\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\right)\mathrm{tr}\left(H_2 - H_1\right) \\
&= \frac{p_1 - p_2}{2}\log(1-\kappa) + \frac{1}{2}\mathrm{tr}\left(\mu_*^T H_1\mu_*\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(\mu_*^T H_2\mu_*\Sigma^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\right)\kappa\left(p_2 - p_1\right)
\end{aligned}$$

Similarly to the proof of the univariate case in Theorem 1, we can relate the terms $\mathrm{tr}\left(\mu_*^T H_i\mu_*\Sigma^{-1}\right)$ to the Kullback-Leibler divergence between $p(Y|\mu_*)$ and $p(Y|\hat{\mu}_i)$. The Kullback-Leibler divergence of $p(Y|\hat{\mu}_i)$ from $p(Y|\mu_*)$ can be derived by employing the usual KL divergence between multivariate normals in (34) to $p(\mathrm{vec}(Y)|\mathrm{vec}(\mu_*))$ and $p(\mathrm{vec}(Y)|\mathrm{vec}(\hat{\mu}_i))$ to obtain

$$\begin{aligned}
\mathrm{KL}(M_*||M_i) &= \frac{1}{2}\left(\log\frac{|\Sigma_i\otimes I_n|}{|\Sigma_*\otimes I_n|} - nd + \mathrm{tr}\left((\Sigma_i^{-1}\otimes I_n)(\Sigma_*\otimes I_n)\right) + (\mathrm{vec}(\mu_*) - \mathrm{vec}(\hat{\mu}_i))^T(\Sigma_i^{-1}\otimes I_n)(\mathrm{vec}(\mu_*) - \mathrm{vec}(\hat{\mu}_i))\right) \\
&= \frac{1}{2}\left(n\log\frac{|\Sigma_i|}{|\Sigma_*|} - nd + n\mathrm{tr}\left(\Sigma_i^{-1}\Sigma_*\right) + \mathrm{tr}\left((\mu_* - \hat{\mu}_i)^T(\mu_* - \hat{\mu}_i)\Sigma_i^{-1}\right)\right) \\
&= \frac{1}{2}\left(-n\log\frac{|\Sigma_*|}{|\Sigma_i|} - nd + n\mathrm{tr}\left(\Sigma_i^{-1}\Sigma_*\right) + \left\|(\mu_* - \hat{\mu}_i)\Sigma_i^{-1/2}\right\|_F^2\right),
\end{aligned}$$

where $\left\|(\mu_* - \hat{\mu}_i)\Sigma_i^{-1/2}\right\|_F^2$ is the Mahalanobis distance with respect to Σ_i. Analogous calculations as in the univariate case gives

$$\mathrm{tr}\left(\mu_*^T H_i\mu_*\Sigma^{-1}\right) = \frac{\mathrm{tr}\left(\mu_*^T\mu_*\Sigma^{-1}\right) - \left\|(\mu_* - \hat{\mu}_i)\Sigma_i^{-1/2}\right\|_F^2}{2-\kappa}$$

and hence we can express the mean as

$$\begin{aligned}
\mathrm{E}(\log \mathrm{B}_{12}(Y)) &= \frac{p_1 - p_2}{2}\log(1-\kappa) + \frac{\frac{1}{2}\left\|(\mu_* - \hat{\mu}_2)\Sigma_2^{-1/2}\right\|_F^2 - \frac{1}{2}\left\|(\mu_* - \hat{\mu}_1)\Sigma_1^{-1/2}\right\|_F^2}{2-\kappa} - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\right)\kappa\left(p_2 - p_1\right) \\
&= \frac{\mathrm{KL}(M_*||M_2) - \mathrm{KL}(M_*||M_1)}{2-\kappa} + \frac{p_1 - p_2}{2}\left(\log(1-\kappa) + \kappa\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\right)\right).
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\log \mathrm{B}_{12}(Y)) &= \frac{1}{2}\mathrm{tr}\left\{\left(\Sigma^{-1}\otimes\left(H_2 - H_1\right)\right)\left(\Sigma_*\otimes I_n\right)\right\}^2 + \mathrm{vec}(\mu_*)^T\left(\Sigma^{-1}\otimes\left(H_2 - H_1\right)\right)\left(\Sigma_*\otimes I_n\right)\left(\Sigma^{-1}\otimes\left(H_2 - H_1\right)\right)\mathrm{vec}(\mu_*) \\
&= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\Sigma^{-1}\Sigma_*\otimes\left(H_2 - H_1\right)^2\right) + \mathrm{vec}(\mu_*)^T\left(\Sigma^{-1}\Sigma_*\Sigma^{-1}\otimes\left(H_2 - H_1\right)^2\right)\mathrm{vec}(\mu_*) \\
&= \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\Sigma_*\Sigma^{-1}\Sigma_*\right)\|H_2 - H_1\|_F^2 + \mathrm{tr}\left(\mu_*^T(H_2 - H_1)^2\mu_*\Sigma^{-1}\Sigma_*\Sigma^{-1}\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\Omega^2\right)\|H_2 - H_1\|_F^2 + \mathrm{tr}\left(\left((\hat{\mu}_2 - \hat{\mu}_1)\Sigma^{-1/2}\right)^T\left((\hat{\mu}_2 - \hat{\mu}_1)\Sigma^{-1/2}\right)\Omega\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\Omega^2\right)\|H_2 - H_1\|_F^2 + \left\|(\hat{\mu}_2 - \hat{\mu}_1)\Sigma^{-1/2}\Omega^{1/2}\right\|_F^2
\end{aligned}$$

where $\Omega \equiv \Sigma^{-1/2}\Sigma_*\Sigma^{-1/2}$ is a multivariate generalization of the variance ratio σ_*^2/σ^2.

Proof of Lemma 1. We first state the following definition from (35).

Definition 1. *Let $S_1, S_2 \subset \mathbb{R}^n$ be subspaces with $p_1 = \dim(S_1) \geq \dim(S_2) = p_2 \geq 1$. The principal angles $\theta_k \in [0, \pi/2]$ between S_1 and S_2 are recursively defined for $k = 1, ..., p_2$ by*

$$\cos(\theta_k) = \max_{u \in S_1} \max_{v \in S_2} |u^T v| = u_k^T v_k, ||u|| = ||v|| = 1,$$

subject to the constraints

$$u_i^T u = 0, \;\; v_i^T v = 0, \;\; i = 1, ..., k-1.$$

Now,

$$\mathrm{tr}(H_1 - H_2)^2 = \kappa^2 \mathrm{tr}(P_1 - P_2)^2 = \kappa^2 \big(\mathrm{tr}(P_1) + \mathrm{tr}(P_2) - 2\mathrm{tr}(P_1 P_2)\big) = \kappa^2 \big(p_1 + p_2 - 2\sum_{i=1}^{n} \lambda_i\big),$$

where λ_i are the eigenvalues of $P_1 P_2$. Theorem 34 in (35) proves that $\mathrm{eig}(P_1 P_2) = (1_s, \cos^2 \theta_{s+i} \; (i = 1, ..., r), 0_{n-s-r})$ where 1_s is an $s \times 1$ vector of ones, 0_{n-s-r} is an $(n - s - r) \times 1$ vector of zeroes, $s = \dim(S_1 \cap S_2)$ is the number of θ_j which are exactly zero, and r is the number of θ_j in the open interval $(0, \pi/2)$. Hence

$$\mathrm{tr}(P_1 - P_2)^2 = \kappa^2 \left(p_1 + p_2 - 2\big(s + \sum_{i=1}^{r} \cos^2(\theta_{k+i})\big) \right).$$