Bayesian Analysis (2019) **14**, Number 3, pp. 703–725

Model Criticism in Latent Space

Sohan Seth[*¶,‖], Iain Murray[†¶], and Christopher K. I. Williams[‡,§¶,‖,**]

Abstract. Model criticism is usually carried out by assessing if replicated data generated under the fitted model looks similar to the observed data, see e.g. Gelman, Carlin, Stern, and Rubin (2004, p. 165). This paper presents a method for latent variable models by pulling back the data into the space of latent variables, and carrying out model criticism in that space. Making use of a model's structure enables a more direct assessment of the assumptions made in the prior and likelihood. We demonstrate the method with examples of model criticism in latent space applied to factor analysis, linear dynamical systems and Gaussian processes.

Keywords: model criticism, latent variable models, factor analysis, linear dynamical systems, Gaussian processes.

1 Introduction

Model criticism is the process of assessing the goodness of fit between some data and a statistical model of that data. Following O'Hagan (2003, p. 423) we prefer the term *model criticism* over *model validation* and *model checking* as it is impossible to validate a model if "all models are wrong", and model criticism has a more active tone of looking to discover problems, compared to model checking, which may seem a more passive activity that does not expect to uncover any problems. While model criticism uses goodness-of-fit tests to judge aspects of the model, its general objective is to identify deficiencies in the model that can lead to *model extension* to address these deficiencies. The extended model(s) can again be subjected to criticism, and the process continues until a *satisfactory* model is found (O'Hagan, 2003). Model criticism is contrasted with *model comparison* in that model criticism assesses a single model, while model comparison deals with at least two models to decide which model is a better fit. Model comparison can be applied to compare the original and the extended model after model criticism and extension (O'Hagan, 2003, p. 2).

Bayesian modelling has become an indispensable tool in statistical learning, and it is being widely used to model complex signals, e.g., by Ratmann et al. (2009). With its growing popularity, there is need for model criticism in this framework. Most work on model criticism makes use of the idea that "if the model fits, then replicated data generated under the model should look similar to observed data" (Gelman et al., 2004,

[*] School of Informatics, University of Edinburgh, UK, sseth@inf.ed.ac.uk

[†] School of Informatics, University of Edinburgh, UK, i.murray@ed.ac.uk

[‡] School of Informatics, University of Edinburgh, UK, ckiw@inf.ed.ac.uk

[§] Alan Turing Institute, London, UK

[¶] Code used in the paper is available at `https://github.com/sohanseth/mcls`.

[‖] SS and CW gratefully acknowledge the UK Engineering and Physical Sciences Research Council (EP/K03197X/1) for funding this work.

[**] The work of CW is supported by EPSRC grant EP/N510129/1 to the Alan Turing Institute.

https://doi.org/10.1214/18-BA1124

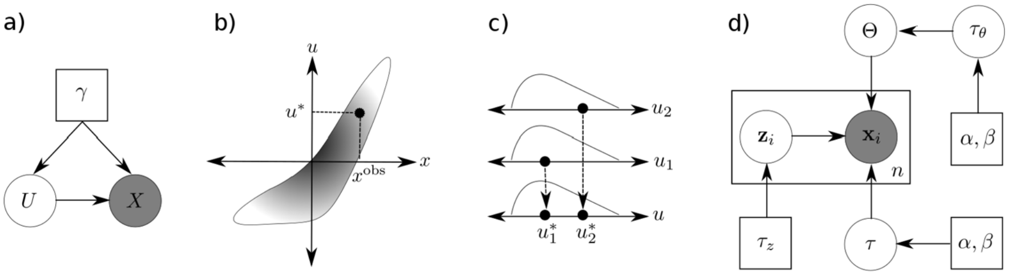


Figure 1: **a)** A probabilistic model with observed variables X, unobserved variables U, and known parameters γ, **b)** Given the observed data x^{obs} from $P(X \,|\, \gamma)$ and a posterior sample u^* from $P(U \,|\, x^{\text{obs}}, \gamma)$, (x^{obs}, u^*) is a joint sample from $P(X, U \,|\, \gamma)$, and therefore, u^* is a sample from $P(U \,|\, \gamma)$ and x^{obs} is a sample from $P(X \,|\, u^*, \gamma)$. **c)** If the prior (or part of it) factorizes into identical distributions, e.g., $P(U \,|\, \gamma) = \prod_{k=1}^{2} P_u(U_k \,|\, \gamma)$, then posterior sample $\{u_1^*, u_2^*\}$ is independent and identical sample from $P_u(\cdot \,|\, \gamma)$. **d)** A factor analysis model showing observed variables $X = \{\mathbf{x}_i\}_{i=1}^n$, unobserved variables $U = \{\{\mathbf{z}_i\}_{i=1}^n, \Theta, \tau, \tau_\theta\}$, and known parameters $\gamma = \{\alpha, \beta, \tau_z\}$. We test if $\{z_{11}^*, z_{12}^*, \ldots\}$ is a sample from $P(Z \,|\, \tau_z)$, and $\{\theta_{11}^*, \theta_{12}^*, \ldots\}$ is a sample from $P(\theta \,|\, \tau_\theta^*)$.

p. 165). In contrast, in this paper we focus on a less well explored idea that for latent variable models, we can probabilistically pull back the data into the space of the latent variables, and carry out model criticism in that space. We can summarize this principle as that *if the model fits, then posterior inferences should match the prior assumptions.*

To elaborate, consider a model with observed variables X and unobserved variables U with joint distribution $P(X, U \,|\, \gamma)$ where γ are known parameters. In general U may contain latent variables Z, parameters Θ, and hyperparameters λ. For example, in the context of the Bayesian matrix factorization (Salakhutdinov and Mnih, 2008), X is the observed data matrix, $U = \{Z, \Theta, \lambda\}$ is the matrix of latent factors Z, the loading matrix Θ, precision hyperparameters λ, and γ denotes the parameters of the hyperpriors. Given a sample x^{obs} from the marginal distribution $P(X \,|\, \gamma)$, and a single posterior sample u^* from the conditional distribution $P(U \,|\, x^{\text{obs}}, \gamma)$, the joint sample (x^{obs}, u^*) is a draw from the distribution $P(X, U \,|\, \gamma)$. This property can be used to check the fit of the model in the latent space by checking if u^* is a sample from the marginal distribution $P(U \,|\, \gamma)$. Testing a single sample against a distribution, however, is not an effective approach. But, in many widely-used models, groups of unknown variables are independently and identically distributed under the prior. These related variables are easily *aggregated* together, giving a simple test of the prior assumptions. Figure 1 summarizes the overall approach, which is justified in §3.

In comparison to model criticism in the observation space, comparing u^* with prior $P(U \,|\, \gamma)$, provides an additional tool for model criticism which does not require crafting an appropriate discrepancy measure, generating replicate observations, and approximating the null distribution. This approach also does not suffer from the "double use" of data (see discussion in §2). These points have also been made by Yuan and Johnson

(2012), but were applied to a relatively small scale hierarchical linear model. We develop the use of model criticism in latent space for large scale and complex models, yielding new insights and developments. Specifically, we apply this approach to the criticism of linear dynamical systems, factor analysis and Gaussian processes, and discuss its connection to the observation space based approach.

The structure of the rest of the paper is as follows: in §2 we describe the methods of model criticism in observation space. §3 provides details of the argument for model criticism in latent space and describes related work, and §4 shows results from applying the method to the three examples. Table 1 describes the notations followed in the paper.

Style	Explanation	Example
Upper case italics	Random variable or a group of random variables	$X, Z, U = \{U_1, \dots, U_K\}$
Lower case italics	Realization of a random variable	$\{x_i\}_{i=1}^n, u^*$
Lower case bold	Vectors, realization or random variable	$\{\mathbf{x}_i\}_{i=1}^n, P(\mathbf{z}), \mathbf{u} = (u_1, \dots, u_K)^\top$
Upper case bold	Matrices, realization or random variable	$\{\mathbf{X}_i\}_{i=1}^n, P(\mathbf{Z}), \mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_K]$
$P(X)$	Distribution of random variable X	$X \sim P(X)$
$P(X \mid y)$	Conditional distribution of r. v. X given $Y = y$	$X \sim P(X \mid y)$
$p(x)$	Probability density function of r. v. X, abbreviation for $p_X(x)$	$p(y) = \int p(x, y) \mathrm{d}x$
$p(x \mid y)$	Conditional density function of r. v. X given $Y = y$, abbreviation for $p_{X \mid Y}(x \mid y)$	$p(y) = \int p(x \mid y) p(y) \mathrm{d}x$
$\cdot \sim \cdot$	Distributed as	$X^{\text{rep}} \sim P(X), X \sim \mathcal{N}(0, 1)$
$\cdot^*$	A posterior sample	$u^*, \mathbf{z}^*$
$\cdot^{\text{obs}}$	Observed data	$X^{\text{obs}} \sim P(X)$
$\cdot^{\text{rep}}$	Replicate data	$X^{\text{rep}} \sim P(X)$
$\mathbf{0}$	Zero vector	$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{I}$	Identity matrix	$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{1}(\cdot)$	Indicator function	$\mathbf{1}(x > y)$

Table 1: Description of notation used in the paper.

2 Model criticism in observation space

A general approach of model criticism is to evaluate if replicated data generated under the (fitted) model looks similar to observed data. Consider that we are modelling observed data x^{obs} with a latent variable model parameterized by U, i.e., we have defined the likelihood $p(x \mid u)$ and (optionally) a prior distribution $P(U)$ over potential parameter values. The principle of model criticism in the observation space is to assess if x^{obs} is a reasonable observation under the proposed model. For example, given the *maximum likelihood estimator* (or another point estimate) $\hat{u}$ of the parameters, one standard approach is to find the *plug-in* p-value (Bayarri and Berger, 2000)

$$p_{\text{plug-in}} = \Pr(D(X^{\text{rep}}, \hat{u}) > D(x^{\text{obs}}, \hat{u})). \tag{1}$$

Here D is called a *discrepancy function* and it resembles a test statistic in hypothesis testing, i.e., a larger value rejects the null hypothesis or indicates incompatibility of data and model, and X^{rep} is a *replicate observation* generated under the fitted model, i.e., $X^{\text{rep}} \sim P(X \mid \hat{u})$.

If the p-value is low, then it implies that the probability of generating a more extreme dataset than the observed data is small, or in other words, the observed data itself is considered extreme relative to the model, and thus, the model does not adequately describe the dataset. In summary, *a low p-value rejects the hypothesis that the data is being adequately modelled*. The p-value is usually estimated via an empirical average by generating multiple replicates x_r^{rep}, $r = 1, \dots, R$, and evaluating

$$\hat{p}_{\text{plug-in}} = \frac{1}{R} \sum_r \mathbf{1}(D(x_r^{\text{rep}}, \hat{u}) > D(x^{\text{obs}}, \hat{u})), \tag{2}$$

where x_r^{rep} is a sample from $P(X \mid \hat{u})$.

An alternative to point estimation is to consider a Bayesian treatment of the problem where one can integrate out the contribution of the parameters. The test statistic can be averaged under either the prior distribution or the posterior distribution. The *prior-predictive* distribution is defined to have the density $p(x^{\text{rep}} \mid \gamma) = \int p(x^{\text{rep}} \mid u)\, p(u \mid \gamma)\, \mathrm{d}u$ where γ parameterizes the prior distribution over U. One can generate replicate observations from this distribution, and compute the *prior predictive p-value* (Box, 1980)

$$\begin{aligned} p_{\text{prior}} &= \Pr(D(X^{\text{rep}}, U) > D(x^{\text{obs}}, U)) \\ &\approx \frac{1}{R} \sum_r \mathbf{1}(D(x_r^{\text{rep}}, u_r) > D(x^{\text{obs}}, u_r)) = \hat{p}_{\text{prior}}, \end{aligned} \tag{3}$$

where $(x^{\text{rep}}, u)_r$ is a sample from $P(X, U \mid \gamma)$. This approach is not reasonable when the prior distribution is improper (cannot be integrated) or uninformative. Additionally, even if the prior distribution is informative, one might not generate enough samples to represent the data distribution well when the parameter space is large. However, notice that one does not need to fit the model to criticise it.

On the other hand, one can use the posterior distribution $P(U \mid x^{\text{obs}})$, and sample from the posterior-predictive distribution with density $p(x^{\text{rep}} \mid x^{\text{obs}}) = \int p(x^{\text{rep}} \mid u)\, p(u \mid x^{\text{obs}})\, \mathrm{d}u$. The *posterior predictive p-value* (Rubin, 1984) is then computed as:

$$\begin{aligned} p_{\text{post}} &= \Pr(D(X^{\text{rep}}, U) > D(x^{\text{obs}}, U) \mid x^{\text{obs}}) \\ &\approx \frac{1}{R} \sum_r \mathbf{1}(D(x_r^{\text{rep}}, u_r) > D(x^{\text{obs}}, u_r)) = \hat{p}_{\text{post}}, \end{aligned} \tag{4}$$

where $(x^{\text{rep}}, u)_r$ is a sample from $P(X^{\text{rep}}, U \mid x^{\text{obs}})$, i.e., by generating samples u_r from the posterior distribution instead. The support of the posterior is usually more concentrated than prior, and the posterior distribution may be well-defined even if the prior distribution is improper. Note that the p-value $p_{\text{post}}(u) = P(D(X^{\text{rep}}, u) > (x^{\text{obs}}, u))$ might be available in closed form depending on the choice of D (Gelman et al., 1996, Eq. (8–9)). Then $p_{\text{post}} = \frac{1}{R} \sum_r p_{\text{post}}(u_r)$ where u_r are posterior samples.

The posterior predictive p-value has been criticised for "double use" of data, once for computing the posterior distribution $P(X^{\text{rep}} \mid x^{\text{obs}})$ and once for computing the discrepancy measure $D(x^{\text{obs}}, U)$ (Bayarri and Berger, 2000). This means that p_{post} does not have a uniform distribution under the null hypothesis, whereas p_{prior} is a valid p-value. $p_{\text{plug-in}}$ is subject to the same criticism as p_{post} since the maximum likelihood estimate (MLE) uses the observed data as well (Bayarri and Berger, 2000). Lloyd and Ghahramani (2015, §7) view the different p-values as arising from "different null hypotheses and interpretations of the word 'model'". They argued that the posterior predictive and plug-in p-values are most useful for highly flexible models, as the aim is to assess the fitted model rather than the whole space of models. Lloyd and Ghahramani (2015) also point out that "it may be more appropriate to hold out data and attempt to falsify the null hypothesis that future data will be generated by the plug-in or posterior distribution", which is also in line with the discussion in (O'Hagan, 2003, §2.1). Further examples of posterior predictive checking can be found in (Belin and Rubin, 1995; Gopalan et al., 2015).

In all of the model criticism described above, a key quantity is the discrepancy function D used to compare the data and predictive simulations. We agree with Belin and Rubin (1995, p. 753) who wrote of the importance of identifying discrepancy functions "that would not automatically be well fit by the assumed model", and that "there is no unique method of Bayesian model monitoring, as there are an unlimited number of non-sufficient statistics that could be studied".

Lloyd and Ghahramani (2015) suggest the Maximum Mean Discrepancy (MMD) as a measure of discrepancy between the observed data and replicates. The motivation of using this approach is to maximize the discrepancy over a class of discrepancy functions rather than choosing only one, i.e.,

$$\text{MMD} = \sup_{f \in \mathcal{F}} (\mathbb{E}_{X^{\text{obs}}} f(X^{\text{obs}}) - \mathbb{E}_{X^{\text{rep}}} f(X^{\text{rep}})). \tag{5}$$

where $\mathcal{F}$ is a set of functions. The function that maximizes the discrepancy is known as the *witness function*. When $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) the witness function can be derived in closed form as

$$\hat{f}(\cdot) = \frac{1}{|x^{\text{obs}}|} \sum_{i=1}^{|x^{\text{obs}}|} \kappa(\cdot, x_i^{\text{obs}}) - \frac{1}{|x^{\text{rep}}|} \sum_{j=1}^{|x^{\text{rep}}|} \kappa(\cdot, x_j^{\text{rep}}), \tag{6}$$

where κ is the kernel of the RKHS. This estimation does not work well in high dimensions, and therefore, the authors suggests reducing the dimensionality of the observation space before applying this statistic (Lloyd and Ghahramani, 2015, p. 4).

3 Model criticism in latent space

Recall we have a model $P(X, U \mid \gamma)$, with observed variables X, unobserved variables U, and known parameters γ. In general U may contain latent variables Z, parameters Θ, and hyperparameters λ. Our procedure depends on the following two key observations:

1. If x^{obs} is drawn from the above model, then a sample u^* from $P(U \mid x^{\text{obs}}, \gamma)$ is a sample from the prior distribution $P(U \mid \gamma)$. To see why this is true, observe that

a natural way to sample from the joint $P(U, X \mid \gamma)$ is to generate a sample u from $P(U \mid \gamma)$, and then generate a sample x from $P(X \mid u, \gamma)$ in that order. However, it is also valid to draw samples from the joint by first sampling x from $P(X \mid \gamma)$ and then sampling u from $P(U \mid x, \gamma)$. Thus we have

Statement 1. *If x^{obs} is a sample from $P(X \mid \gamma)$, then a sample u^* from $P(U \mid x^{\mathrm{obs}}, \gamma)$ will be a draw from $P(U \mid \gamma)$.*

It is important to clarify what Statement 1 is *not* saying. It is not saying that repeated draws from $P(U \mid x^{\mathrm{obs}}, \gamma)$ will explore the full prior distribution $P(U \mid \gamma)$, but only that it is a valid way to draw *one* sample from it if x^{obs} is a draw from the model. However, testing how well a single draw from a given distribution fits that distribution is difficult. This brings us to our second observation.

2. If U is a collection of variables, i.e., $U = (U_1, \ldots, U_K)$, and the prior distribution of U decomposes into independent draws from the same distribution, e.g., $P(U \mid \gamma) = \prod_{k=1}^{K} P_u(U_k \mid \gamma)$ then it is possible to *aggregate* these variables together, i.e., instead of testing if $(u_1^*, \ldots, u_K^*)$ is a sample from $P(U \mid \gamma)$, one can test if $\{u_1^*, \ldots, u_K^*\}$ is independent and identical draws from the distribution $P_u(\cdot \mid \gamma)$. In other words, rather than testing one sample against a known high dimensional distribution, one can test if the collection of K samples are independent and identical draws from a known lower-dimensional distribution P_u. Thus, we define aggregation as *pooling variables with the same prior distribution together*, and an *aggregated posterior sample* (APS) is defined as a set of posterior samples that have been aggregated for comparison with a specific *reference distribution*. The above can be generalized to the situation where $U = (U_1, \ldots, U_K, \theta)$ is a collection of variables and parameters such that $P(U \mid \gamma) = \prod_{k=1}^{K} P_u(U_k \mid \theta) P(\theta \mid \gamma)$. Then $\{u_1^*, \ldots, u_K^*\}$ can be aggregated and tested against $P_u(\cdot \mid \theta^*)$. Alternatively, U and θ can be combined to define a *pivotal quantity* s whose distribution does not depend on θ (Yuan and Johnson, 2012), and $s(u^*, \theta^*)$ can be tested against that distribution. Aggregation can be also extended to the case where U consists of groups of variables $(U^1, \ldots, U^G)$ where aggregation is performed within each group $U^g = (U_1^g, \ldots, U_{K_g}^g)$ by pooling $\{u_1^{g*}, \ldots, u_{K_g}^{g*}\}$ and comparing against $p_{u_g}(\cdot \mid u^{-g*})$ where U^{-g} denotes all groups except g. We provide more concrete examples of aggregation in §3.1 and Table 2.

We refer to this approach as *aggregated posterior checking* (APC). We summarize this approach in Algorithm 1. We assume that the prior distribution is proper, so the respective posterior distribution is well-defined, and that any Markov chain Monte Carlo (MCMC) sampler has converged, i.e., the posterior sample is well-behaved. Ideas equivalent to Statement 1 and the aggregation of posterior samples can also be found in Yuan and Johnson (2012)[1], but were applied to the case where U contains only model parameters, and for hierarchical linear models. See §3.2 for more details on related work.

[1] We had independently derived the key results. We thank an anonymous referee for pointing out the work of Yuan and Johnson (2012).

Model	x^{obs}	U	APS	reference distribution
MF (13)	$\{\mathbf{x}_i^{\text{obs}}\}_{i=1}^n$	$\{\mathbf{z}_i\}_{i=1}^n, \boldsymbol{\Theta}, \mathbf{b}, \tau, \tau_z$ for (14)-(15) $\{\mathbf{z}_i\}_{i=1}^n, \boldsymbol{\Theta}, \mathbf{b}, \tau, \boldsymbol{\pi}, \boldsymbol{\tau}$ for (16)	$\{z_{ki}^*\}_{i=1,k=1}^{n,K}$, and $\{(z_{k_1 i}^*, z_{k_2 i}^*)\}_{i=1,\, k_1,k_2=1,k_1\neq k_2}^{n,K}$	$\mathcal{N}(0, {\tau_z^*}^{-1})$, and $\mathcal{N}(0, {\tau_z^*}^{-1}\mathbf{I}_2)$ for (14) $\mathcal{L}(0, \tau_z^*)$, and $\mathcal{L}(z_1; 0, \tau_z^*)\, \mathcal{L}(z_2; 0, \tau_z^*)$ for (15) $\sum_m \pi_m^* \mathcal{N}(0, {\tau_m^*}^{-1})$, and $\sum_m \pi_m^* \mathcal{N}(\mathbf{0}, {\tau_m^*}^{-1}\mathbf{I}_2)$ for (16)
LDS (17), (18)	$\{(x, y,$ $\cos(\nu), \sin(\nu))_t^{\text{obs}}\}_{t=1}^n$	$\{s_t\}_{t=1}^n, \{\mathbf{z}_t\}_{t=1}^n, \mathbf{A}^{(1)}, \ldots,$ $\mathbf{A}^{(S)}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(S)}, \mathbf{B}, \mathbf{R}$	$\{\tilde{z}_{kt}^*\}_{t=2}^n$ from (19) and $\{\tilde{x}_{jt}^*\}_{t=2}^n$ from (20) $\forall\, j, k$ $\{(\tilde{z}_{kt}^*, \tilde{z}_{k(t+1)}^*)\}_{t=2}^{n-1}$ from (19) and $\{(\tilde{x}_{kt}^*, \tilde{x}_{k(t+1)}^*)\}_{t=2}^{n-1}$ from (20) $\forall\, j, k$	$\mathcal{N}(0, 1)$ $\mathcal{N}(0, 1)$ $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$
GP (9)	$\{(x_i, y_i)^{\text{obs}}\}_{i=1}^n$	σ_f^2, l, τ for (22) σ_f^2, p, l_p, l_d for (22) and (23) $\sigma_f^2, f, l_p, l_d, \sigma_{fs}^2, l_s, \sigma_{fl}^2, l_l$ for (22) (large and small) and (23)	$\{z_i^*\}_{i=1}^n$ from (21)	$\mathcal{N}(0, 1)$

Table 2: The table summarizes observed data x^{obs}, unknown variables U, aggregated posterior sample(s) (APS), and corresponding reference distribution(s) (as elaborated in Algorithm 1) for three models discussed in §4, and different scenarios within each model.

Algorithm 1 Aggregated posterior check

```
Require: Observed data $x^{\mathrm{obs}}$
Require: Bayesian model $P(X \mid U, \gamma)P(U \mid \gamma)$ with latent variables $U$
1: Generate a posterior sample $u^*$ from $P(U \mid x^{\mathrm{obs}}, \gamma)$
2: Generate aggregated posterior sample(s)    ▷ See Table 2
3: Compare aggregated posterior sample(s) with corresponding reference distribution(s) with appropriate test
4: return p-value of the test(s)
```

So far, we have addressed the idea of assessing deviations from the prior distribution and aggregation in the latent space. However, the same idea can be applied to the observation space as well, i.e., to the likelihood by testing if x^{obs} is a sample from $P(X \mid u^*, \gamma)$. Although it is true that a discrepancy in the choice of likelihood should be reflected in the posterior sample, assessing the discrepancy in the likelihood directly provides better understanding and easier resolution of the discrepancy. Notice that, although we make use of x^{obs}, our approach is not equivalent to model criticism in the observation space since we do not compare the observed data x^{obs} with replicate observations x_r^{rep}, but only investigate the relation between the latent space u^* and observation space x^{obs}. Both methods, however, require generating posterior samples u_r (for model criticism in the latent space we use $r = 1$).

3.1 Application to different models

We discuss below the application of model criticism in latent space to factor analysis, linear dynamical systems and Gaussian process regression. These situations are then demonstrated on real data in §4.

Factor analysis model Consider a factor analysis model with hyperparameters $\lambda = \{\tau_\theta, \tau\}$, parameters (loading matrix) $\mathbf{\Theta}$, latent variables (factors) $\mathbf{z}$ and data $\mathbf{x}$. Grouping $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$ and similarly for $\mathbf{X}$,

$$\begin{aligned} p(\lambda, \mathbf{\Theta}, \mathbf{Z}, \mathbf{X}) &= p(\lambda)\, p(\mathbf{\Theta} \mid \lambda)\, p(\mathbf{Z} \mid \lambda)\, p(\mathbf{X} \mid \mathbf{Z}, \mathbf{\Theta}, \lambda) \\ &= p(\tau_\theta)\, p(\tau) p(\mathbf{\Theta} \mid \tau_\theta) \prod_{i=1}^{n} p(\mathbf{z}_i)\, p(\mathbf{x}_i \mid \mathbf{z}_i, \mathbf{\Theta}, \tau). \end{aligned} \tag{7}$$

Figure 1d illustrates this model. We have omitted the fixed parameters γ for simplicity. In Gaussian factor analysis, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, {\tau_z}^{-1}\mathbf{I})$ and $\boldsymbol{\theta} \mid \lambda \sim \mathcal{N}(\mathbf{0}, {\tau_\theta}^{-1}\mathbf{I})$. There is an identifiability issue in the factor analysis model between $\mathbf{\Theta}$ and $\mathbf{z}$, which is resolved by fixing the scale of one of the two. In (7) the dependence of $\mathbf{z}$ on λ is taken to be null, i.e., $\tau_z = 1$. (In the case of example §4.1 we fix the scale of $\mathbf{\Theta}$ instead.) Also, $P(\mathbf{x} \mid \mathbf{z}, \mathbf{\Theta}, \tau) = \mathcal{N}(\mathbf{\Theta}\mathbf{z}, \tau^{-1}\mathbf{I})$ and $\tau, \tau_\theta \mid \alpha, \beta \sim \mathrm{Gamma}(\alpha, \beta)$. Thus the fixed parameters $\gamma = \{\tau_z, \alpha, \beta\}$.

If $\mathbf{X}^{\mathrm{obs}}$ is drawn from the above model, then a sample $\lambda^*, \mathbf{\Theta}^*, \mathbf{Z}^*$ from $P(\lambda, \mathbf{\Theta}, \mathbf{Z} \mid \mathbf{X}^{\mathrm{obs}})$ is a sample from the prior $P(\lambda)P(\mathbf{\Theta} \mid \lambda)\, P(\mathbf{Z} \mid \lambda)$. In factor analysis, $\mathbf{Z}$ decomposes into independent draws from $P(\mathbf{z}\,|\,\tau_z)$, and therefore, one can pool the posterior samples $\mathbf{z}_i^*$ to assess deviations from $P(\mathbf{z}\,|\,\tau_z)$. Moreover, each $\mathbf{z}_i$ usually decomposes into independent draws over the different latent dimensions as $\prod_k p(z_{ik}\,|\,\tau_z)$, one can pool the z_{ik}^* to assess deviations from $p(z\,|\,\tau_z)$. Similarly, if the prior over the factor loadings matrix $\boldsymbol{\Theta}$ decomposes as $p(\boldsymbol{\Theta}\,|\,\tau_\theta) = \prod_{kj} p(\theta_{kj}\,|\,\tau_\theta)$ then one can pool the θ_{kj}^*s, and compare with $p(\theta\,|\,\tau_\theta^*)$. One can also go beyond the marginal z or the full vector $\mathbf{z}$, and assess a subset of the vector such as bivariate interactions (see §4.1).

Linear dynamical system One can extend the idea of aggregation beyond factor analysis models. For example, Statement 1 holds for general latent variable models with *repeated structure*. Take, for example, a linear dynamical system model with a latent Markov chain, so that

$$p(\mathbf{X},\mathbf{Z}\,|\,U) = p(\mathbf{z}_1)p(\mathbf{x}_1\,|\,\mathbf{z}_1,U)\prod_{t=2}^{T} p(\mathbf{z}_t\,|\,\mathbf{z}_{t-1},U)p(\mathbf{x}_t\,|\,\mathbf{z}_t,U), \tag{8}$$

where U consists of the system and observation matrices $\mathbf{A},\mathbf{B}$, and precisions, $\mathbf{Q},\mathbf{R}$. Then according to Statement 1 a sample $(\mathbf{Z}^*,u^*)$ drawn from $P(\mathbf{Z},U\,|\,\mathbf{X}^{\text{obs}})$ should be distributed according to the prior over $(\mathbf{Z},U)$. Although the $\mathbf{z}_t$'s are not independent (due to the Markov chain), we can consider model criticism for $p(\mathbf{z}_t\,|\,\mathbf{z}_{t-1})$. For example for a system model parameterized as $\mathbf{z}_t\,|\,\mathbf{z}_{t-1} \sim \mathbf{A}\mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t$ with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0},\mathbf{Q}^{-1})$, violations of the model will show up as deviations of the $\boldsymbol{\epsilon}_t^*$'s from $\mathcal{N}(\mathbf{0},\mathbf{Q}^{*-1})$ (see §4.2). Similarly, for an observation model parameterized as $\mathbf{x}_t\,|\,\mathbf{z}_t \sim \mathbf{B}\mathbf{z}_t + \boldsymbol{\psi}_t$ with $\boldsymbol{\psi}_t \sim \mathcal{N}(\mathbf{0},\mathbf{R}^{-1})$, violations of the model may also show up as deviations of the $\boldsymbol{\psi}_t^*$'s from $\mathcal{N}(\mathbf{0},\mathbf{R}^{*-1})$. See §4.2.

Gaussian process regression A Gaussian process probabilistic model is defined as:

$$\vartheta,\zeta,\tau \sim p(\vartheta)\,p(\zeta)\,p(\tau), \tag{9a}$$

$$f(x) \sim \mathcal{GP}(m(x\,|\,\vartheta),\,\kappa(x,x'\,|\,\zeta)), \tag{9b}$$

$$y_i \sim \mathcal{N}(f(x_i),\tau^{-1}) \quad \forall i = 1,\ldots,n, \tag{9c}$$

where $m(x\,|\,\vartheta)$ is the mean function parameterized by ϑ, $\kappa(x,x'\,|\,\zeta)$ is the covariance function (or kernel) parameterized by ζ, and τ is the observation noise precision, see e.g., Rasmussen and Williams (2006). Given observations $\{(\mathbf{x}_i,y_i)\}_{i=1}^n$, $\mathbf{y} \sim \mathcal{N}(\mathbf{m},\mathbf{K})$, where $\mathbf{y} = (y(\mathbf{x}_1),\ldots,y(\mathbf{x}_n))^\top$, $\mathbf{m} = (m(\mathbf{x}_1\,|\,\vartheta),\ldots,m(\mathbf{x}_n\,|\,\vartheta))^\top$ and $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j\,|\,\zeta) + \tau^{-1}\delta(\mathbf{x}_i,\mathbf{x}_j)$. Alternatively, considering the eigendecomposition $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^\top$ where $\mathbf{U} = [\mathbf{u}_1,\ldots,\mathbf{u}_n]$ is the matrix of the eigenvectors $\mathbf{u}_i$s and Λ is the diagonal matrix of the corresponding eigenvalues, i.e., $\Lambda_{ii} = \lambda_i$,

$$\mathbf{c} = \mathbf{U}^\top(\mathbf{y}-\mathbf{m}) \sim \mathcal{N}(\mathbf{0},\Lambda). \tag{10}$$

This implies that, according to the model, the projections c_i of the signal $\mathbf{y}$ on the eigenvector $\mathbf{u}_i$ are independent samples from $\mathcal{N}(0,\lambda_i)$. Thus, the normalized projections

$$\mathbf{z} = \Lambda^{-1/2}\mathbf{U}^\top(\mathbf{y}-\mathbf{m}) \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \tag{11}$$

should be independent samples from $\mathcal{N}(0, 1)$ distribution. One can thus test the normality of the z's to assess the goodness of fit. However, note that if the ith eigenvalue of $\mathbf{K}$ is much smaller than the noise variance τ^{-1}, then this z_i is dominated by the white noise contribution. Thus we only include z's corresponding to eigenvalues with $\lambda_i > 2\tau^{-1}$ to assess the fit of the GP model. The factor of 2 on the right-hand side (RHS) of the inequality is included because the λ_i's are shifted by τ^{-1} by definition. See §4.3.

3.2 Related work

Cook et al. (2006) consider the situation with (in our notation) a prior $p(u)$ on the parameters of the model, and likelihood $p(x \mid u)$. They then assume that specific parameters u^0 are drawn from the prior, then data x^{obs} drawn from $P(X \mid u^0)$. They then consider samples $u^1, u^2, \ldots, u^L$ drawn from $P(U \mid x^{\mathrm{obs}})$, and comment (in the caption of their Figure 1, translating to our notation) that "$(x^{\mathrm{obs}}, u^\ell)$ should look like a draw from $P(X, U)$ for $\ell = 0, 1, \ldots, L$". They then use the 'reverse' of Statement 1 to validate the correctness of posterior samples generated by a statistical software, by comparing u^0 with $u^1, u^2, \ldots, u^L$. Their recommended method for this is to calculate posterior quantiles for each scalar parameter; if the software is working correctly then the posterior quantiles are uniformly distributed. Although they share with us the observation that $(x^{\mathrm{obs}}, u^\ell)$ should look like a draw from $P(X, U)$, this is used to answer a totally different question. Also, they do not discuss the inclusion of latent variables in the model.

Johnson (2007) and later Yuan and Johnson (2012) also consider a model with parameters U and data drawn from $P(X \mid U)$. Their interest is in the use of pivotal quantity $d(x, u)$ that has a known and invariant sampling distribution when data x^{obs} are generated from a model with data-generating parameters u^0. Then Yuan and Johnson (2012) show that if the $d(X, u^0)$ is a pivotal quantity distributed according to F, then $d(X, u^\ell)$ is also distributed according to F, if u^ℓ is drawn from the posterior on U given x^{obs}. The result of Yuan and Johnson (2012) extends earlier work by Johnson (2007) to the case where $d(x, u)$ does not depend on the data x—for example this situation can arise in a Bayesian hierarchical linear regression model, when considering the second level where parameters for individual units are generated from a hyperprior.

Regression diagnostics is a well-explored example of model criticism. Existing approaches assess certain statistical assumptions made during modelling, e.g., if the residuals follow a normal distribution with zero mean, (e.g., using a Q-Q plot (Wilk and Gnanadesikan, 1968)), if the residuals are homoscedastic, (e.g., using the Breusch–Pagan (Breusch and Pagan, 1979) or White test (White, 1980)) or if the successive residual terms are uncorrelated (e.g., using the Durbin–Watson test (Durbin and Watson, 1950)). Regression diagnostics can be seen as a special case of model criticism in the latent space since residuals are representatives of *errors*, which are latent variables of the model. However, our methods are also applicable to more complex models.

Meulders et al. (1998) consider a factor analysis model for binary data, using (in our notation) $\mathrm{Beta}(2, 2)$ priors on $\mathbf{Z}$ and $\boldsymbol{\Theta}$. They carry out posterior sampling using block Gibbs sampling for $\mathbf{Z}$ and $\boldsymbol{\Theta}$ and compare histograms of these variables against the prior. Discrepancies between the prior and histograms of the sampled aggregated posterior led to model extension, expanding the model to use a mixture of two beta distributions for the parameters. However, the authors do not explain the basis for carrying out this check (cf. Statement 1).

Buccigrossi and Simoncelli (1999) consider the posterior distribution of wavelet coefficients (analogous to $\mathbf{z}$ in the factor analysis model) in response to image patches. By considering the distribution of a bivariate aggregated posterior, they show that this is not equal to the product of the marginals, but exhibits variance correlations. (This is shown by introducing a "bowtie plot" showing the conditional histogram of z_2 given z_1.) This work is a nice example of how the failure of a diagnostic test can give rise to an extended model (see §4.1).

O'Hagan (2003, §3) considered model criticism tools that can be applied at each node of a graphical model (and of course latent variables can be considered as such). O'Hagan (§3.1 2003) discussed the idea of residual testing at different levels of a hierarchical model as well as a generic probabilistic model. He suggested checking if a node in a probabilistic model is misbehaving by comparing the posterior samples at that node to prior distribution. O'Hagan (§3.2 2003) also emphasized that conflict can arise between the different sources of information about a variable at a particular node, arising from contributions from each neighbouring node in the graph. However, he did not suggest using the aggregated posterior to assess goodness of fit, but considered the posterior at each node separately.

Tang et al. (2012) introduce the concept of the "aggregated posterior" as applied to deep mixtures of factor analysers (MFA) model. Consider the situation as above but where $\mathbf{\Theta}$ is estimated by maximum likelihood, so it is the posterior over Z that is of interest. Thus $p(\mathbf{X}, \mathbf{Z} \,|\, \mathbf{\Theta}) = \prod_{i=1}^{n} p(\mathbf{z}_i)\, p(\mathbf{x}_i \,|\, \mathbf{z}_i, \mathbf{\Theta})$. Under this model we also have that $p(\mathbf{z}) = \int p_{\mathbf{\Theta}}(\mathbf{z} \,|\, \mathbf{x})\, p_{\mathbf{\Theta}}(\mathbf{x}) d\mathbf{x}$ where the $\mathbf{\Theta}$ subscript denotes that both $p_{\mathbf{\Theta}}(\mathbf{z} \,|\, \mathbf{x})$ and $p_{\mathbf{\Theta}}(\mathbf{x})$ correspond to distributions under the model. Tang et al. (2012, p3) define the aggregated posterior as "the empirical average over the data of the posteriors over the factors", i.e.,

$$\tilde{p}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} p_{\mathbf{\Theta}}(\mathbf{z} \,|\, \mathbf{x}_i^{\mathrm{obs}}), \tag{12}$$

where the integral with respect to $p_{\mathbf{\Theta}}(\mathbf{x})$ has been replaced by the empirical average over samples. If the data distribution $p(\mathbf{x})$ is equal to the model distribution $p_{\mathbf{\Theta}}(\mathbf{x})$ then $\tilde{p}(\mathbf{z})$ should agree with $p(\mathbf{z})$. However, differences between $p(\mathbf{x})$ and $p_{\mathbf{\Theta}}(\mathbf{x})$ will manifest as differences between the two respective distributions in the latent space.

In practice, however, one does not explicitly construct the aggregated posterior (12) since it is only asymptotically equal to the prior. Instead Tang et al. (2012) compare a collection of n samples $\mathbf{z}_i^*$ from $p_{\mathbf{\Theta}}(\mathbf{z} \,|\, \mathbf{x}_i^{\mathrm{obs}})$ for $i = 1, \ldots, n$ to $p(\mathbf{z})$. This is a valid approach since if $\{\mathbf{x}_1^{\mathrm{obs}}, \ldots, \mathbf{x}_n^{\mathrm{obs}}\}$ follow the distribution $p_{\mathbf{\Theta}}(\mathbf{x})$, then $\{\mathbf{z}_1^*, \ldots, \mathbf{z}_n^*\}$ follow the distribution $p(\mathbf{z})$ as we show in Statement 1. Additionally, as $\mathbf{\Theta}$ is not known in practice, Tang et al. (2012) replace $\mathbf{\Theta}$ with maximum likelihood estimate $\hat{\mathbf{\Theta}}$ in the definition of aggregated posterior (12). In Statement 1 we extend this idea to a Bayesian setting where $\mathbf{\Theta}$ and λ are not fixed parameters but latent variables themselves.

Tang et al. (2012) started with a simple mixture of factor analysers (MFA), and observed that the aggregated posterior for a latent component often doesn't match the $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior. By replacing the prior for a component with another MFA model, they constructed a deep MFA model. The idea of the aggregated posterior (although not the name) can be traced back e.g. to Hinton et al. (2006), where in deep belief nets the idea was that the posterior distribution of the latents of a restricted Boltzmann machine (RBM) could be modelled by another RBM.

4 Examples

In this section, we provide three examples of model criticism and extension in the latent space. First, we explore a factor analysis model in the context of image compression (§4.1). The objective of this example is to show how the model can be criticised in the latent space as well as in the observation space. Our analysis leads to changing the latent distribution from a single Gaussian to a scale mixture of Gaussians, which captures both the marginal and the joint structure of the latent space, and improves the model in the observation space as well.

Next, we explore a linear dynamical system model (§4.2) in the context of modelling time series. We show that model criticism in latent space allows us to interrogate not only the standard "innovations" (defined in (20)), but also the latent residuals (defined in (19)).

Finally, we explore a Gaussian process model (§4.3) in the context of modelling time series. The objective of this example is to show when model criticism in the latent space can be a natural choice whereas model criticism in the observation space can be difficult. Our analysis leads to changing the covariance function from squared exponential to a combination of periodic and squared exponential kernels.

We implemented all models (except the Gaussian process model) in JAGS (Plummer, 2003), keeping a single sample in the MCMC run after discarding a burn-in of 1000 samples (10,000 samples for §4.2). Note that for model criticism in the latent space, we need *only a single sample*. We summarize the aggregation process and corresponding reference distributions used in this section in Table 2.

4.1 Image patch data

The Berkeley Segmentation Database (Martin et al., 2001) consists of 200 training images. Following Zoran and Weiss (2012), we convert the images to greyscale, extract 8 × 8 randomly located patches, and remove the mean from all image patches[2]. We extract 50,000 image patches, and fit different matrix factorization models of the form:

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \tau \sim \mathrm{Gamma}(\alpha, \beta),\ \theta_{jk} \sim \mathcal{N}(0, 1) \tag{13a}$$

$$\mathbf{z}_i \sim \mathrm{LatentDist},\ \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\Theta}\mathbf{z}_i + \mathbf{b},\ \tau^{-1}\mathbf{I}) \tag{13b}$$

[2] https://people.csail.mit.edu/danielzoran/NIPSGMM.zip

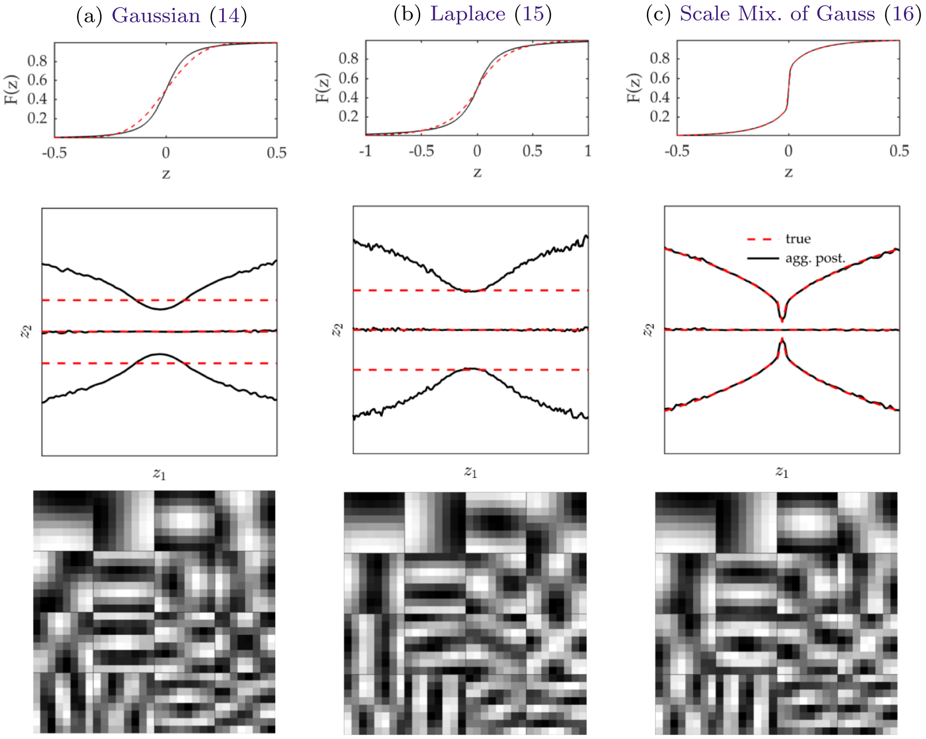


Figure 2: Aggregated posterior [agg. post] and associated prior at the latent node for (left to right) the Gaussian, Laplace and Scale mixture of Gaussian models of image patches. **Top**: Empirical cumulative distribution function (ECDF) of aggregated posterior samples ($16 \times 50,000$ samples) and cumulative distribution function (CDF) of respective prior. **Middle**: conditional mean and standard deviation of the bivariate aggregated posterior samples ($16 \times 15 \times 50,000 \div 2$ samples, over 100 bins) and respective prior distribution. **Bottom**: eigenvectors of respective loading matrices.

with $K = 16$ latent dimensions ($> 82\%$ explained variance in PCA). Previously Zoran and Weiss (2012) used full covariance zero-mean Gaussians ($\tau = 0$, $K = 64$, $\mathbf{b} = \mathbf{0}$). We set $\alpha = \beta = 0.001$.

We start by assuming a Gaussian distribution for the latent model

$$\tau_z \sim \text{Gamma}(\alpha, \beta),\; z \sim \mathcal{N}(0, {\tau_z}^{-1}), \qquad \text{(14, Gaussian)}$$

and generate a sample $(\mathbf{Z}^*, \mathbf{\Theta}^*, \mathbf{b}^*, \tau^*, \tau_z^*)$ from the posterior. To criticise the model, we aggregate the univariate posterior samples $\{z_{ki}^*\}\, \forall k, i$, and bivariate samples $\{(z_{ki_1}^*, z_{ki_2}^*)\}$ $\forall k, i_1 \neq i_2$. If the observed data follows the model, then the distributions of the corresponding APSs are univariate normal $\mathcal{N}(0, {\tau_z^*}^{-1})$ and bivariate normal $\mathcal{N}(\mathbf{0}, {\tau_z^*}^{-1}\mathbf{I}_2)$ respectively. We observe that neither of the APSs follow the expected prior distribution. Also, the marginal distribution is more concentrated around zero than the expected distribution, whereas the joint distribution shows heteroscedasticity (Figure 2, left column) which is inconsistent with the factorized bivariate normal prior.

An alternative latent variable model for the factor analysis model is

$$z \sim \mathcal{L}(0, \tau_z). \tag{15, Laplace}$$

We use the same aggregation strategy as before, and compare the empirical distributions with the univariate distribution $\mathcal{L}(0, \tau_z^*)$ and bivariate distribution $\mathcal{L}(z_1; 0, \tau_z^*)\mathcal{L}(z_2; 0, \tau_z^*)$ respectively. We observe similar characteristics in the aggregated posterior as before (Figure 2, middle column).

To accommodate these observations, we allow a scale mixture of Gaussian distributions as used by Wainwright and Simoncelli (2000) with 8 components for the latent variable

$$\pi \sim \mathrm{Dir}(\mathbf{1}),\ \tau_m \sim \mathrm{Gamma}(\alpha, \beta),\ \mathbf{z} \sim \sum_{m=1}^{8} \pi_m \mathcal{N}(\mathbf{0}, {\tau_m}^{-1}\mathbf{I}). \tag{16, Scale Mix. of Gauss}$$

We generate sample $(\pi_1^*, \ldots, \pi_8^*, \tau_1^*, \ldots, \tau_8^*)$ as well. We assess the same aggregated distributions as before and compare them with $\sum_m \pi_m^* \mathcal{N}(0, {\tau_m^*}^{-1})$, and $\sum_m \pi_m^* \mathcal{N}(\mathbf{0}, {\tau_m^*}^{-1}\mathbf{I}_2)$ respectively. We observe that the empirical marginal distribution follows the mixture distribution well, although a KS test rejects the hypothesis that the aggregated posterior follows the mixture distribution. Additionally, the joint distribution captures the heteroscedasticity in the latent space (Figure 2, right column).

We also show the eigenvectors of the corresponding loading matrix for each of the three cases (Figure 2 bottom row). We show the eigenvectors rather than the loading matrix themselves since for the Gaussian and Gaussian scale mixture, the columns of the loading matrix may not correspond to any particular pattern due to rotational invariance. We observe that all three loading matrices span a similar space.

The matrix factorization model can be criticised in the observation space with established image statistics as a discrepancy measure. This, however, requires generating replicate data of the same size as the observed data, which in this case is computationally extensive since $\mathbf{X}^{\mathrm{obs}} \in \mathbb{R}^{64\times 50{,}000}$. To avoid generating multiple replicates, i.e., matrices $\mathbf{X}_r^{\mathrm{rep}} \in \mathbb{R}^{64\times 50{,}000}$, we only generate a single replicate for each latent distribution choice and compare them the observed data.

For all three cases, i.e., Gaussian, Laplace, and Scale Mix. of Gauss, we generate latent samples $\mathbf{z}_i^{\mathrm{rep}}$ from the fitted parameters τ_z^* (and $\boldsymbol{\tau}, \boldsymbol{\pi}^*$ for Scale Mix. of Gauss). We use the rest of the fitted parameters, i.e., $\boldsymbol{\Theta}^*$, τ^*, and b^* to generate samples $\mathbf{x}_i^{\mathrm{rep}}$ from $\mathbf{z}_i^{\mathrm{rep}}$. We generate 50,000, 8×8 replicate image patches, and compare the observed and replicate data in terms of the distribution of raw pixel values. We show the results in Figure 3. We observe that the distribution of the image pixel values in the replicate data follows the observed data more closely for Scale Mix. of Gauss than the other latent distributions. However, it is not a perfect fit, and that tells us that this model can improved further; potentially by increasing K, and varying the noise characteristics such as using a full diagonal covariance.

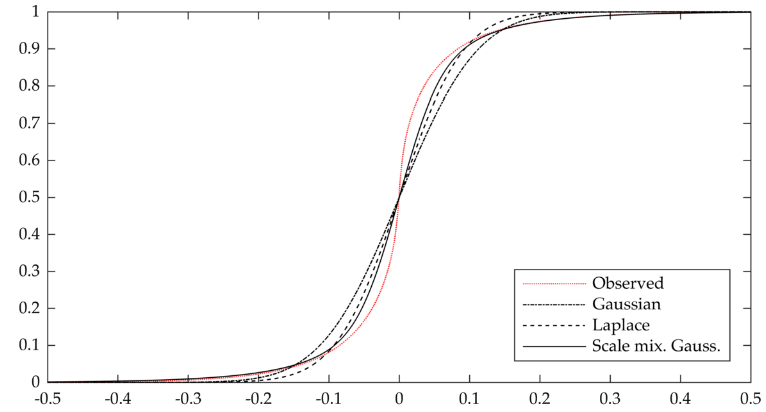


Figure 3: Empirical cumulative distribution functions of raw pixel values on observed and replicate data for varying different distributions.

4.2 Honey bee data

The honey bee data consists of measurements of (x, y) coordinate and head angle (ν) of 6 honey bees. The measurements are usually translated into a 4-dimensional multivariate time series $(x, y, \cos(\nu), \sin(\nu))$, and modelled using a switching linear dynamical system to capture three distinct dynamical regimes, namely, left turn, right turn and waggle (Oh et al., 2008). We follow this strategy, and model each time series by a switching linear dynamical system (SLDS) (Fox et al., 2009) as follows:

$$s_1 = 1,\ \mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{17a}$$

$$s_t \sim \mathrm{Cat}(\boldsymbol{\pi}^{(s_{t-1})}) \qquad \forall t = 2, \ldots, n \tag{17b}$$

$$\mathbf{z}_t \sim \mathbf{A}^{(s_t)}\mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{(s_t)^{-1}}) \qquad \forall t = 2, \ldots, n \tag{17c}$$

$$\mathbf{x}_t \sim B\mathbf{z}_t + \boldsymbol{\psi}_t,\ \boldsymbol{\psi}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{-1}) \qquad \forall t = 1, \ldots, n, \tag{17d}$$

where s_t can be in one of $\{1, \ldots, S\}$ states. We assume that $\mathbf{Q}^{(\cdot)}$ (for each state) and $\mathbf{R}$ are diagonal matrices with Gamma(α, β) prior over nonzero entries, entries of $\mathbf{A}^{(\cdot)}$ (for each state) and $\mathbf{B}$ originate from Gaussian distribution, and $\boldsymbol{\pi}^{(\cdot)}$ (for each state) follow a Dirichlet distribution, i.e.,

$$a_{\cdot\cdot}^{(\cdot)} \sim \mathcal{N}(0, {\tau_A}^{-1}),\ b_{\cdot\cdot} \sim \mathcal{N}(0, {\tau_B}^{-1}) \tag{18a}$$

$$\tau_A \sim \mathrm{Gamma}(\alpha, \beta),\ \tau_B \sim \mathrm{Gamma}(\alpha, \beta) \tag{18b}$$

$$\boldsymbol{\pi}^{(\cdot)} \sim \mathrm{Dir}(\mathbf{1}). \tag{18c}$$

We group $\mathbf{s} = \{s_i\}_{i=1}^n$ and $Z = \{\mathbf{z}_i\}_{i=1}^n$. We set $\alpha = \beta = 0.001$.

We fit two models with $S = 1$ (standard linear dynamical system), and $S = 3$, both with a 4 dimensional latent space. We generate a posterior sample $(\mathbf{s}^*, \mathbf{Z}^*, \mathbf{A}^{(1)*},$

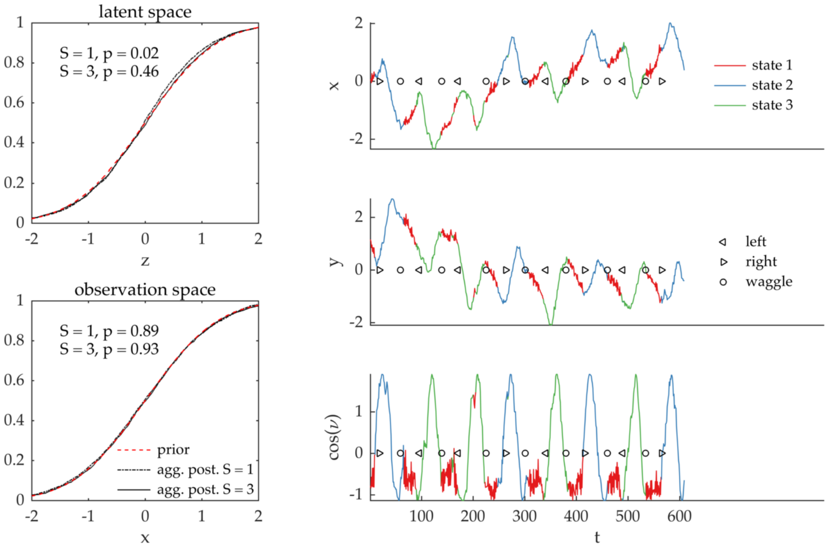


Figure 4: **Left**: ECDF of aggregated posterior samples [agg. post.] ($4 \times n$ samples) for $S = 1$ and $S = 3$, and CDF of prior $\mathcal{N}(0, 1)$ at the latent (top) and observation node (bottom). **Right**: segmentation of honey bee sequence 6 as observed in $s_1^*, \dots, s_n^*$. Black markers indicate true change points. p-values correspond to KS test.

$\dots, \mathbf{A}^{(S)*}, \mathbf{Q}^{(1)*}, \dots, \mathbf{Q}^{(S)*}, \mathbf{B}^*, \mathbf{R}^*)$, and aggregate the *standardized* latent residuals

$$\tilde{\mathbf{z}}_t = (\mathbf{Q}^{(s_t^*)*})^{0.5}(\mathbf{z}_t^* - \mathbf{A}^{(s_t^*)*}\mathbf{z}_{t-1}^*) \,\forall\, t = 2, \dots, n, \tag{19}$$

and observation residuals (or innovations)

$$\tilde{\mathbf{x}}_t = (\mathbf{R}^*)^{0.5}(\mathbf{x}_t^{\text{obs}} - \mathbf{B}^*\mathbf{z}_t^*) \,\forall\, t = 2, \dots, n. \tag{20}$$

For linear dynamical system (LDS) the standard approach to model criticism is to check that the innovations sequence is zero-mean and white (see, e.g. (Candy, 1986, §5.1)), although this is usually carried out for known or point-estimates of the parameters, not in a Bayesian setting. We use this check (extended to the SLDS case) below, but also consider the latent residuals.

First, we focus on marginal structures $\tilde{z}_{kt}$ and $\tilde{x}_{jt}$ by pooling $k = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ together, rather than the 4-dimensional vectors themselves as shown in Figure 4 (left). We expect that the APSs would deviate from normality more (lower p-value) when $S = 1$, compared to $S = 3$, and we observe this to be true for all honey bee sequences except 2. For sequences 4–6, the latent segmentations of the SLDS in terms of $(s_1^*, \dots, s_n^*)$ agree with the ground truth well; we present the 6th sequence in Figure 4 (right). For sequences 1–3, we observe that the segmentations are rather poor, similar to the results in (Fox et al., 2009, §5).

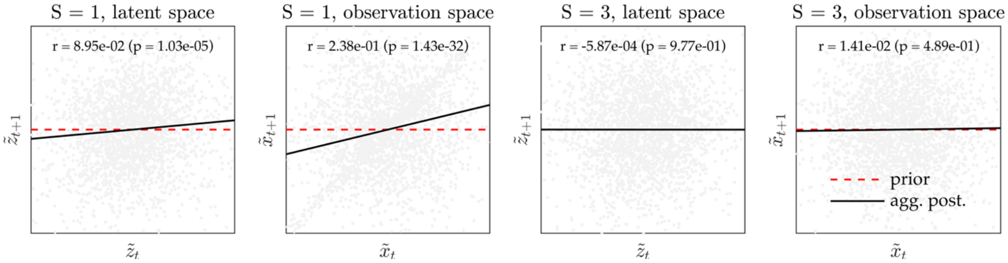


Figure 5: Scatterplot of bivariate aggregated posterior samples (607 × 4 samples) for sequence 6. The black line shows the best linear fit, whereas the dotted line shows the expected fit in the absence of (serial) correlation. The values in each plot are the correlation (r) and respective p-value (p).

Next, we focus on the joint structures in the temporal domain by pooling, $(\tilde{x}_{j_1 t}, \tilde{x}_{j_2 t})$ $\forall j_1, j_2 = 1, 2, 3, 4$ and $j_1 \neq j_2$, $(\tilde{z}_{kt}, \tilde{z}_{k(t+1)})$ $\forall k = 1, 2, 3, 4$, and $(\tilde{x}_{jt}, \tilde{x}_{j(t+1)})$ $\forall j = 1, 2, 3, 4$. We expect that this APSs would deviate from the reference distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ more for $S = 1$ than for $S = 3$. We compute the correlation coefficients, and observe the respective p-values for $S = 1$ and $S = 3$ (Rahman, 1968). We observe that for $S = 1$, the models are rejected either in the latent domain or in the observation domain except for sequence 3, while for $S = 3$, the models are rejected either in the latent domain or the observation domain for sequences 2 and 3 only. In other words, for sequences 1, 4, 5 and 6, the model improves for $S = 3$, whereas for sequence 2 it fails to improve, and for sequence 3 it degrades for $S = 3$. These observations can again be attributed to the poor segmentation for sequences 1-3. We show the corresponding aggregated posteriors in the latent and observation space for sequence 6 in Figure 5. We observe that the residuals in both latent and observation space display correlations for $S = 1$ while these is reduced considerably for $S = 3$.

4.3 Carbon emission data

The CO2 emission dataset[3] comprises monthly average atmospheric carbon concentration y_i (in parts per million) between 1958 and 2017 (707 measurements after removing missing values). Rasmussen and Williams (2006, §5.4.3) show that this time series can be modelled well by a combination of 4 standard covariance functions involving 10 hyperparameters (and an additional parameter to model the additive white noise). Each covariance function is introduced to model a specific aspect of the signal, e.g., a squared exponential kernel to model the long term trend, a decaying periodic kernel to model the seasonal variation, a rational quadratic kernel to model the short term irregularities, and another squared exponential kernel to model the residual correlated noise. We show below how model criticism and extension can be used to justify the use of covariance functions representing similar aspects of the data.

[3] ftp://ftp.cmdl.noaa.gov/ccg/co2/trends/co2_mm_mlo.txt

Following Rasmussen[4] we use the measurements up to year 2004 (543) as training data and the rest (164) as testing data. Notice that we do not use testing data for model criticism but to show the goodness of fit visually. We remove the mean of the training data before modelling, and use a zero mean function, i.e., $m(x) = 0$ or $\mathbf{m} = \mathbf{0}$. We use Gamma(α, β) priors over ζ and τ^{-1} to keep them positive. We use the GPstuff toolbox (Vanhatalo et al., 2013) to generate MCMC samples, and initialize the sampler at the maximum likelihood (ML) solution obtained using Gaussian Processes for Machine Learning (GPML) toolbox[5]. We set the parameters α and β such that the mean of the prior distribution is at the ML solution, and the variance is equal to the mean. We generate a posterior sample $(\vartheta^*, \zeta^*, \tau^*\}$ and aggregate the standardized projections

$$\mathbf{z}^* = {\Lambda^*}^{-1/2} {\mathbf{U}^*}^\top (\mathbf{y} - \mathbf{m}^*), \tag{21}$$

where $\mathbf{U}^*$ and Λ^* are the eigenvectors and eigenvalues of kernel matrix $\mathbf{K}^*$ such that $\mathbf{K}^*_{ij} = \kappa(x_i, x_j \,|\, \zeta^*) + {\tau^*}^{-1} \delta(x_i, x_j)$, and $\mathbf{m}^* = 0$ by design.

We first model the time series with the squared exponential or Gaussian kernel,

$$\kappa_{\mathrm{se}}(x, x' \,|\, \zeta) = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right), \tag{22}$$

where l is the length scale and σ_f^2 is the signal variance, i.e., $\zeta = \{\sigma_f^2, l\}$. We obtain $\zeta_{\mathrm{ML}} = (188, 0.30)$ and $\zeta^* = (197, 0.29)$. It is also possible to model this time series with a large length scale, i.e., $\zeta = (1958, 31)$ but this has lower marginal likelihood $\exp(-1198)$ as opposed to $\exp(-753)$. We present the fitted data along with unstandardized and standardized projections in Figure 6. The figure shows that the Gaussian kernel fails to model the time series as the prediction quickly falls to the mean of the training signal and KS-test p-value $= 4 \times 10^{-10}$. We observe that most of the signal strength ($\mathbf{c}_i$'s) is concentrated at lower frequencies (corresponding to large eigenvalues $\lambda_{i \in \{1,5\}}$). The respective eigenvectors correspond to an upward trend. Also, a relatively high strength is observed at eigenvalues 92–93. The respective eigenvectors correspond to sinusoids of frequency ~1 year (see Figure 7a) which indicates a potential need of a periodic covariance function to model this data.

To tackle this, we use the decaying periodic function to model the time series, (Rasmussen and Williams, 2006, §5.4.3)

$$\kappa_{\mathrm{pe}}(x, x' \,|\, \zeta) = \sigma_f^2 \exp\left(-\frac{2\sin^2(\pi(x - x')/p)}{l_p^2}\right) \exp\left(-\frac{(x - x')^2}{2l_d^2}\right), \tag{23}$$

where p is the period of the covariance function. Therefore, $\zeta = (\sigma_f^2, p, l_p, l_d)$. We obtain $\zeta_{\mathrm{ML}} = (283, 1, 5.13, 5.86)$, and $\zeta^* = (385, 1, 4.88, 6.09)$. We observe that this provides a better fit than squared exponential kernel (p-value 0.08). Although the KS-test fails to reject the fitted model (perhaps due to lack of samples), we observe that the signal strengths ($\mathbf{c}_i$'s) still deviate from their expected values. In particular, the second, fourth

[4] http://learning.eng.cam.ac.uk/carl/mauna/

[5] http://www.gaussianprocess.org/gpml/code/matlab/doc/

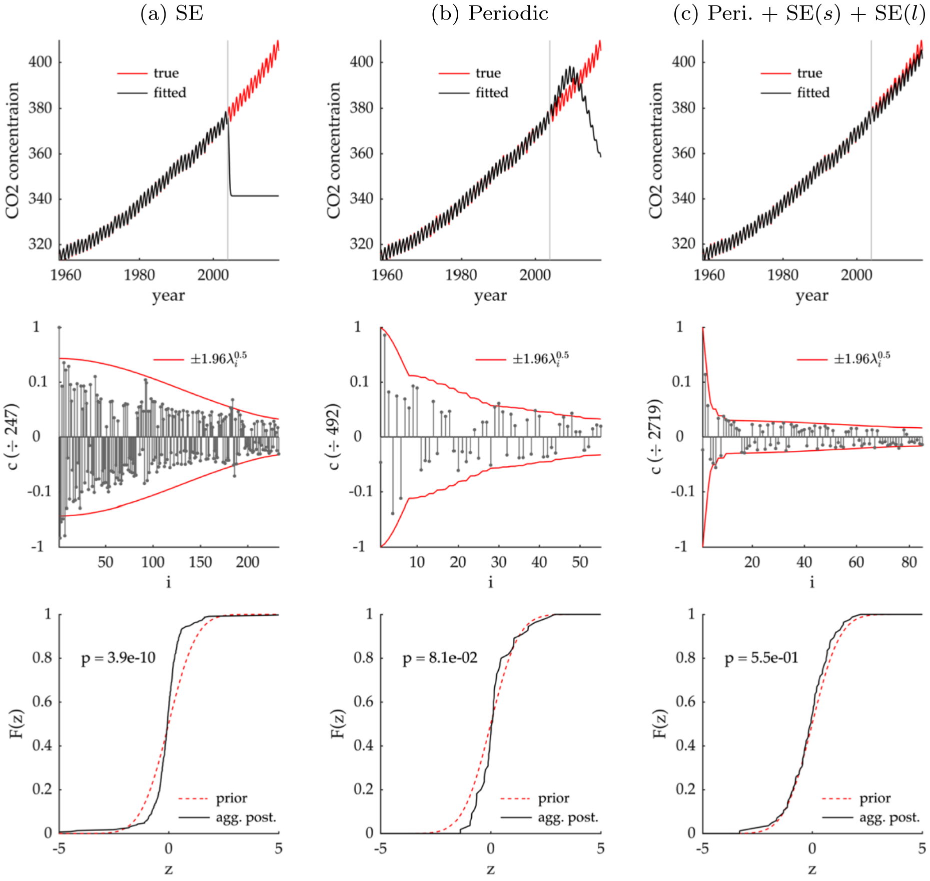
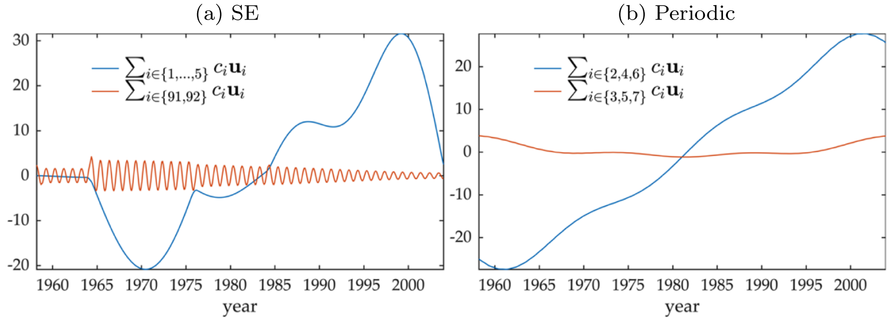


Figure 6: Latent values of the Gaussian process model for CO2 emission dataset. **Top**: Original and fitted signal. Training and testing sets are separated by a gray line. **Middle**: Unnormalized projections $\mathbf{c}_i^*$'s for $i = 1, \ldots, n$, and the respective 95% confidence interval $\pm 1.96\lambda_i^{1/2}$. We only show values for which $\lambda_i^* > 2\tau^{*-1}$. *y-axis has been transformed by* $\operatorname{sgn}(y)|y|^{0.3}$ *to show small values.* **Bottom**: ECDFs of aggregated posterior samples [agg. post.] of the normalized projections $\mathbf{z}_i$'s and CDF of prior distribution $\mathcal{N}(0, 1)$. p-values correspond to KS-test.

and sixth projections show relatively high values compared to third, fifth and seventh. The signal $\sum_{i\in\{2,4,6\}} c_i\mathbf{u}_i$ corresponds to an upward trend, which corroborates the need to model the trend further. See Figure 7b. Note that although the CO2 data is a time-series, the analysis of the **c**-samples (see (10)) does not depend on this, and can also be used where the input-space is multi-dimensional.

To accommodate the upward trend, we introduce a squared exponential kernel with a relatively large length scale. However, to avoid modelling small scale variations with the same kernel, we use combination of two squared exponential kernels with two different


(a) SE

(b) Periodic

Figure 7: Weighted eigenfunctions of carbon emission dataset. The kinks in the plot appear due to the short length scale.

length scales. Therefore, $\zeta = (\sigma_f^2, p, l_p, l_d, \sigma_{fs}^2, l_s, \sigma_{fl}^2, l_l)$ where the last four parameters belong to the two squared exponential kernels with small (s) and large (l) length scales. We obtain $\zeta_{\mathrm{ML}} = (4.37, 1, 1.78, 74.60, 0.81, 0.92, 4132, 27.14)$, and $\zeta_* = (2.25, 1, 1.24, 73.88, 0.32, 0.66, 4095, 32.25)$. We observe that this improves the fit even further, both in terms of the testing data (visually) and in terms of unstandardized projections. The KS-test uses more samples, and still fails to reject the model (p-value 0.55).

Model criticism of Gaussian processes in the observation space has been discussed by Lloyd and Ghahramani (2015). However, their approach is different from the standard posterior predictive check since the authors use hold-out data rather than using the observed data twice. Although this approach shows if the response on hold-out data is different for the fitted model, it does not necessarily point out how the model can be extended.

One could generate a replicate sample from $\mathbf{y}^{\mathrm{rep}} \sim P(\cdot \mid \zeta^*, \mathbf{X}^{\mathrm{obs}})$, and compare $\mathbf{y}^{\mathrm{rep}}$ and $\mathbf{y}^{\mathrm{obs}}$ as for a posterior predictive check. However, note that in this case $\mathbf{y}^{\mathrm{rep}}$ will be an *independent* draw from the GP with parameters ζ^* and input locations $\mathbf{X}^{\mathrm{obs}}$, hence it could look very different from $\mathbf{y}^{\mathrm{obs}}$—this is why Lloyd and Ghahramani (2015) make use of held-out data. Also, it would be difficult to come up with a suitable discrepancy function in this case. One could consider the χ^2 discrepancy[6], i.e., $\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}$. However, this quantity is fitted when sampling the kernel parameters ζ, and is also (as discussed above) dominated by the noise for small eigenvalues of $\mathbf{K}$. Other discrepancy measures could be investigated, but exploring these alternatives is beyond the scope of this paper.

5 Discussion

Model criticism explores the discrepancies between a statistical model $P(X, U)$ and observed data x^{obs}. This is often achieved by generating replicate observations $X^{\mathrm{rep}} \sim P(X \mid x^{\mathrm{obs}})$ from the fitted model, and investigating which aspects $D(X, U)$ of the replicated observations do not match the observed data. Instead here we have focused on

[6] Inspired by Gelman et al. (1996, Eq. (8))

pulling the effect of the data back into the latent space, and investigating if the posterior sample $u^* \sim P(U \,|\, x^{\mathrm{obs}})$ follows the prior distribution $P(U)$, as it should do by Statement 1 if the data were generated by the model. This is tested by aggregating related variables with the same prior distribution and comparing them with the associated prior.

It should be noted that model criticism is not used to judge if a model is right or wrong. On the contrary, it is widely accepted that *all models are wrong but some are useful* (Box and Draper, 1987, p. 424). Model criticism aims at understanding the limitations of the model with the hope that a better model can be found, e.g., since all models are basically simplifications of a more complex process, model criticism inspects if the simplification is meaningful, or if the statistical assumptions made are reasonable. Following this principle, we have discussed four examples of model criticism in latent space. We have shown that by analysing the distribution of the aggregated posterior, a model can be extended so that the aggregated posteriors follow the respective prior distributions better.

References

Bayarri, M. J. and Berger, J. O. (2000). "p-values for Composite Null Models." *Journal of the American Statistical Association*, 95(452): 1127–1142. MR1804239. doi: https://doi.org/10.2307/2669749. 705, 706, 707

Belin, T. R. and Rubin, D. B. (1995). "The Analysis of Repeated-Measures Data on Schizophrenic Reaction Times using Mixture Models." *Statistics in Medicine*, 14(8): 747–768. 707

Box, G. E. (1980). "Sampling and Bayes' Inference in Scientific Modelling and Robustness." *Journal of the Royal Statistical Society*, 143(4): 383–430. MR0603745. doi: https://doi.org/10.2307/2982063. 706

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley. MR0861118. 723

Breusch, T. S. and Pagan, A. R. (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*, 47(5): 1287–1294. MR0545960. doi: https://doi.org/10.2307/1911963. 712

Buccigrossi, R. P. and Simoncelli, E. P. (1999). "Image Compression via Joint Statistical Characterization in the Wavelet Domain." *IEEE Transactions on Signal Processing.*, 8(12): 1688–1701. 713

Candy, J. V. (1986). *Signal Processing: The Model Based Approach*. McGraw-Hill. 718

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). "Validation of Software for Bayesian Models Using Posterior Quantiles." *Journal of Computational and Graphical Statistics*, 15(3): 675–692. MR2291268. doi: https://doi.org/10.1198/106186006X136976. 712

Durbin, J. and Watson, G. S. (1950). "Testing for Serial Correlation in Least Squares Regression: I." *Biometrika*, 37(3/4): 409–428. MR0039210. doi: `https://doi.org/10.1093/biomet/37.3-4.409`. 712

Fox, E., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2009). "Nonparametric Bayesian Learning of Switching Linear Dynamical Systems." In *Advances in Neural Information Processing Systems 21*, 457–464. 717, 718

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. London: Chapman and Hall. Second edition. MR2027492. 703

Gelman, A., Meng, X., and Stern, H. (1996). "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies." *Statistica Sinica*, 733–807. MR1422404. 706, 722

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). "Scalable Recommendation with Hierarchical Poisson Factorization." In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI*, 326–335. 707

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation*, 18: 1527–1554. MR2224485. doi: `https://doi.org/10.1162/neco.2006.18.7.1527`. 714

Johnson, V. E. (2007). "Bayesian model assessment using pivotal quantities." *Bayesian Analysis*, 2(4): 719–733. MR2361972. doi: `https://doi.org/10.1214/07-BA229`. 712, 725

Lloyd, J. R. and Ghahramani, Z. (2015). "Statistical Model Criticism Using Kernel Two Sample Tests." In *Advances in Neural Information Processing Systems*. 707, 722

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics." In *Proceedings of 8th International Conference on Computer Vision*, volume 2, 416–423. 714

Meulders, M., Gelman, A., Van Mechelen, I., and De Boeck, P. (1998). "Generalizing the Probability Matrix Decomposition Model: an Example of Bayesian Model Checking and Model Expansion." In Hox, J. J. and de Leeuw, E. D. (ed.), *Assumptions, Robustness and Estimation Methods in Multivariate Modeling*. Amsterdam: TT-Publikaties. 712

Oh, S. M., Rehg, J. M., Balch, T., and Dellaert, F. (2008). "Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems." *International Journal of Computer Vision*, 77(1): 103–124. 717

O'Hagan, A. (2003). "HSSS Model Criticism." In Green, P. J., Hjort, N. L., and Richardson, S. (eds.), *Highly Structured Stochastic Systems*, 422–444. Oxford University Press. MR2082418. 703, 707, 713

Plummer, M. (2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. 714

Rahman, N. A. (1968). *A Course in Theoretical Statistics*. Charles Griffin and Company. 719

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. MR2514435. 711, 719, 720

Ratmann, O., Andrieu, C., Wiuf, C., and Richardson, S. (2009). "Model criticism based on likelihood-free inference, with an application to protein network evolution." *Proceedings of the National Academy of Sciences*, 106(26): 10576–10581. 703

Rubin, D. B. (1984). "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *Annals of Statistics*, 12: 1151–1172. MR0760681. doi: https://doi.org/10.1214/aos/1176346785. 706

Salakhutdinov, R. and Mnih, A. (2008). "Bayesian Probabilistic Matrix Factorization using Markov chain Monte Carlo." In *Proceedings of the International Conference on Machine Learning*, volume 25. 704

Tang, Y., Salakhutdinov, R., and Hinton, G. E. (2012). "Deep Mixtures of Factor Analysers." In *Proceedings of the 29th International Conference on Machine Learning*. 713

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). "GPstuff: Bayesian Modeling with Gaussian Processes." *Journal of Machine Learning Research*, 14(1): 1175–1179. MR3063621. 720

Wainwright, M. J. and Simoncelli, E. P. (2000). "Scale Mixtures of Gaussians and the Statistics of Natural Images." In *Advances in Neural Information Processing Systems*, volume 12, 855–861. 716

White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4): 817–838. MR0575027. doi: https://doi.org/10.2307/1912934. 712

Wilk, M. B. and Gnanadesikan, R. (1968). "Probability Plotting Methods for the Analysis of Data." *Biometrika*, 55(1): 1–17. 712

Yuan, Y. and Johnson, V. E. (2012). "Goodness-of-fit diagnostics for Bayesian hierarchical models." *Biometrics*, 68(1): 156–164. MR2909864. doi: https://doi.org/10.1111/j.1541-0420.2011.01668.x. 704, 708, 712, 725

Zoran, D. and Weiss, Y. (2012). "Natural Images, Gaussian Mixtures and Dead Leaves." In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, 1736–1744. Curran Associates, Inc. 714, 715

Acknowledgments

We thank the anonymous referees for pointing out the work of Johnson (2007) and Yuan and Johnson (2012), and for comments which helped to improve the paper.