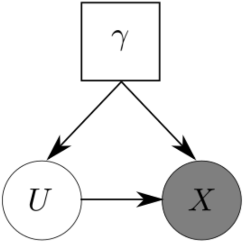
Model Criticism in Latent Space

Sohan Seth

University of Edinburgh, School of Informatics, Edinburgh, EH8 9AB, UK

September 10, 2019

THE UNIVERSITY *of* EDINBURGH



Estimate the mean μ of n observations

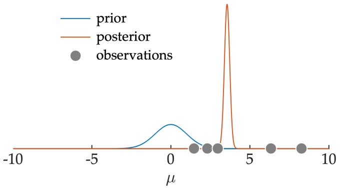
$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) && \text{(prior)} \\ X_i \mid \mu &\sim \mathcal{N}(\mu, \sigma^2) && \text{(likelihood)} \end{aligned}$$

- X denotes observed variables
- U denotes unknown variables
- γ denotes known variables

- $X = \{X_1, X_2, \ldots, X_n\}$
- $U = \{\mu\}$
- $\gamma = \{\mu_0, \sigma_0^2, \sigma^2\}$

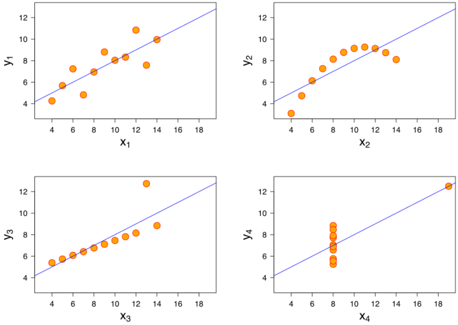
Bayes' rule

$$p_{U \mid X}(u \mid x^{\text{obs}}, \gamma) = \frac{p_{X \mid U}(x^{\text{obs}} \mid u, \gamma) p_U(u \mid \gamma)}{p_X(x^{\text{obs}} \mid \gamma)}$$



Motivation

- Statistical models are approximation of complex natural processes
- "all models are wrong but some are useful" [Box and Draper, 1987, p. 424]
- Is the simplification meaningful?
- Are the assumptions we make reasonable?
- Knowing the limitations can guide us to build a better model
- *Model criticism* is the process of assessing the limitations of a model



[SOURCE]

Figure: Anscombe's Quartet [Francis Anscombe 1973]

Model Criticism in Observation Space

"if the model fits, then replicated data [X^{rep}] generated under the model should look similar to observed data [in terms of discrepancy measure D]" [Gelman et al., 2004, p. 165]

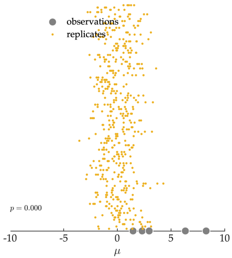
$$D(x, u) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

"if the model fits, then replicated data [X^{rep}] generated under the model should look similar to observed data [in terms of discrepancy measure D]" [Gelman et al., 2004, p. 165]

- *prior predictive p-value* [Box, 1980]

$$p_{\text{prior}} = \Pr(D(X^{\text{rep}}, U) > D(x^{\text{obs}}, U)) \text{ where } X^{\text{rep}}, U \sim P(X, U) \tag{1}$$

$$D(x, u) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$



Model Criticism in Observation Space

"if the model fits, then replicated data [X^{rep}] generated under the model should look similar to observed data [in terms of discrepancy measure D]" [Gelman et al., 2004, p. 165]

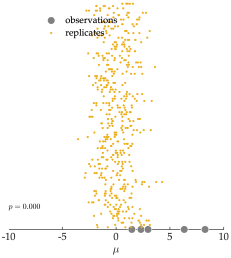
- *prior predictive p-value* [Box, 1980]

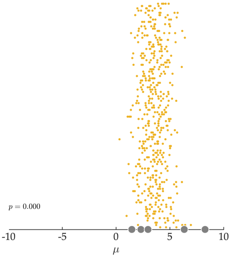
$$p_{\text{prior}} = \Pr(D(X^{\text{rep}}, U) > D(x^{\text{obs}}, U)) \text{ where } X^{\text{rep}}, U \sim P(X, U) \tag{1}$$

- *posterior predictive p-value* [Rubin, 1984]

$$p_{\text{post}} = \Pr(D(X^{\text{rep}}, U) > D(x^{\text{obs}}, U) \mid x^{\text{obs}}) \text{ where } X^{\text{rep}}, U \sim P(X, U \mid x^{\text{obs}}) \tag{2}$$

$$D(x, u) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$





Model Criticism in Latent Space

Posterior predictive check requires [Johnson, 2007]

1. generating replicate observations
2. crafting an appropriate discrepancy measure
3. approximating the null distribution, and
4. "double use" of data

Model Criticism in Latent Space

Posterior predictive check requires [Johnson, 2007]

1. generating replicate observations
2. crafting an appropriate discrepancy measure
3. approximating the null distribution, and
4. "double use" of data

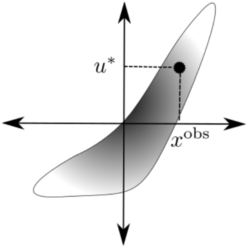
If the model fits, then posterior inferences should match the prior assumptions.

Model Criticism in Latent Space

Posterior predictive check requires [Johnson, 2007]

1. generating replicate observations
2. crafting an appropriate discrepancy measure
3. approximating the null distribution, and
4. "double use" of data

If the model fits, then posterior inferences should match the prior assumptions.



$$x^{\mathrm{obs}} \sim P(X) \text{ and } u^* \mid x^{\mathrm{obs}} \sim P_{U \mid X}(u \mid x^{\mathrm{obs}}) \Rightarrow (u^*, x^{\mathrm{obs}}) \sim P_{U,X}(u, x) \Rightarrow u^* \sim P(U)$$

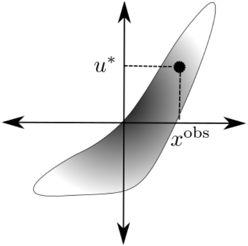
If x^{obs} is a sample from $P(X \mid \gamma)$, then a sample u^* from $P(U \mid x^{\mathrm{obs}}, \gamma)$ will be a draw from $P(U \mid \gamma)$.

Model Criticism in Latent Space

Posterior predictive check requires [Johnson, 2007]

1. generating replicate observations
2. crafting an appropriate discrepancy measure
3. approximating the null distribution, and
4. "double use" of data

If the model fits, then posterior inferences should match the prior assumptions.



$$x^{\text{obs}} \sim P(X) \text{ and } u^* \mid x^{\text{obs}} \sim P_{U \mid X}(u \mid x^{\text{obs}}) \Rightarrow (u^*, x^{\text{obs}}) \sim P_{U,X}(u, x) \Rightarrow u^* \sim P(U)$$

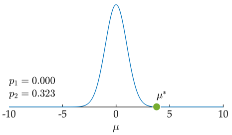
If x^{obs} is a sample from $P(X \mid \gamma)$, then a sample u^* from $P(U \mid x^{\text{obs}}, \gamma)$ will be a draw from $P(U \mid \gamma)$.

$u_1^*, \ldots, u_m^* \not\sim P(U \mid \gamma)$, i.e., m posterior samples are not independent samples from the prior

Aggregated Posterior Check

Require: Observed data x^{obs}
Require: Bayesian model $P(X \mid U, \gamma)P(U \mid \gamma)$ with latent variables U

1: Generate a posterior sample u^* from $P(U \mid x^{\text{obs}}, \gamma)$
2:
3: Compare sample with corresponding prior distribution
4: **return** p-value of the test

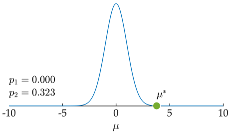


$$(1)\ \mu^* \sim \mathcal{N}(\mu_0, \sigma_0^2) \text{ and } (2)\ X_1, \ldots, X_n \sim \mathcal{N}(\mu^*, \sigma^2) \tag{3}$$

Aggregated Posterior Check

Require: Observed data x^{obs}
Require: Bayesian model $P(X \mid U, \gamma)P(U \mid \gamma)$ with latent variables U

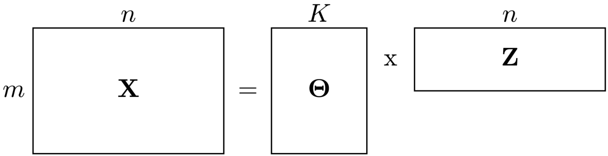
1: Generate a posterior sample u^* from $P(U \mid x^{\text{obs}}, \gamma)$
2: Generate aggregated posterior sample
3: Compare aggregated posterior sample with corresponding prior distribution
4: **return** p-value of the test



$$(1)\ \mu^* \sim \mathcal{N}(\mu_0, \sigma_0^2) \text{ and } (2)\ X_1, \ldots, X_n \sim \mathcal{N}(\mu^*, \sigma^2) \tag{3}$$

- Often U is a collection of variables, i.e., $U = (U_1, \ldots, U_K)$, and $P(U \mid \gamma) = \prod_{k=1}^{K} P_u(U_k \mid \gamma)$
- Instead of testing if $(u_1^*, \ldots, u_K^*)$ is a sample from $P(U \mid \gamma)$, test if the *aggregated* variables $\{u_1^*, \ldots, u_K^*\}$ is independent and identical draws from $P_u(\cdot \mid \gamma)$.

1. Probabilistic Matrix Factorization



For $i = 1, \ldots, n$

$$\mathbf{z}_i \sim \text{LatentDist} \mid \tau_z \tag{4}$$

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\Theta}\mathbf{z}_i + \mathbf{b},\ \tau^{-1}\mathbf{I}) \tag{5}$$

Given $\mathbf{Z}^* = [\mathbf{z}_1^*, \ldots, \mathbf{z}_n^*]$ and τ_z^*, (e.g., $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \tau_z^{-1}\mathbf{I})$)

$$\{z_{ki}^*\} \sim P(z \mid \tau_z^*) \quad \text{(univariate)}$$

$$\{(z_{k_1 i}^*, z_{k_2 i}^*) : k_1 \neq k_2\} \sim P(z_1, z_2 \mid \tau_z^*) \quad \text{(bivariate)}$$

1. Image Patches [Zoran and Weiss, 2012]

$n = 50,000$, 8×8 image patches, i.e., $m = 64$ and we consider $k = 16$

$$\tau_z \sim \text{Gamma}(\alpha, \beta),\ z \sim \mathcal{N}(0, \tau_z{}^{-1}), \tag{6}$$

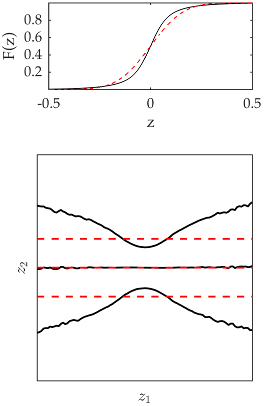


Figure: Dotted line is the prior distribution and straight line is the aggregated posterior

1. Image Patches [Zoran and Weiss, 2012]

$n = 50,000$, 8×8 image patches, i.e., $m = 64$ and we consider $k = 16$

$$\tau_z \sim \text{Gamma}(\alpha, \beta),\ z \sim \text{Laplace}(0, \tau_z). \tag{6}$$

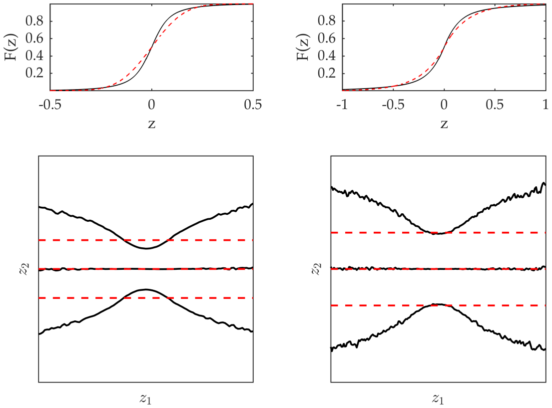


Figure: Dotted line is the prior distribution and straight line is the aggregated posterior

1. Image Patches [Zoran and Weiss, 2012]

$n = 50,000$, 8×8 image patches, i.e., $m = 64$ and we consider $k = 16$

$$\pi \sim \text{Dir}(\mathbf{1}),\ \tau_m \sim \text{Gamma}(\alpha, \beta),\ \mathbf{z} \sim \sum_{m=1}^{8} \pi_m \mathcal{N}(\mathbf{0}, {\tau_m}^{-1}\mathbf{I}) \quad (6)$$

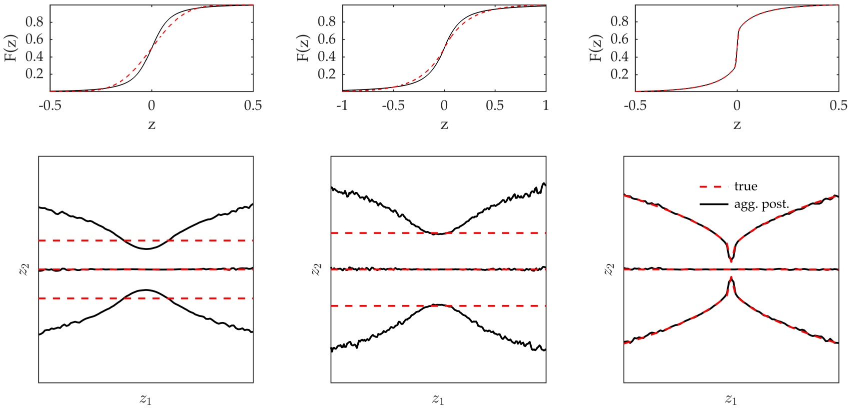
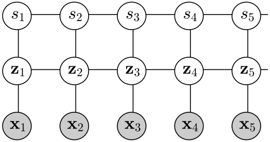


Figure: Dotted line is the prior distribution and straight line is the aggregated posterior



$$s_1 = 1,\ \mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{7}$$

$$s_t \sim \text{Cat}(\boldsymbol{\pi}^{(s_{t-1})}) \qquad \forall t = 2, \ldots, n \tag{8}$$

$$\mathbf{z}_t \sim \mathbf{A}^{(s_t)} \mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{(s_t)\,-1}) \qquad \forall t = 2, \ldots, n \tag{9}$$

$$\mathbf{x}_t \sim B\mathbf{z}_t + \boldsymbol{\psi}_t,\ \boldsymbol{\psi}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{-1}) \qquad \forall t = 1, \ldots, n \tag{10}$$

Standardized residuals follow $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution

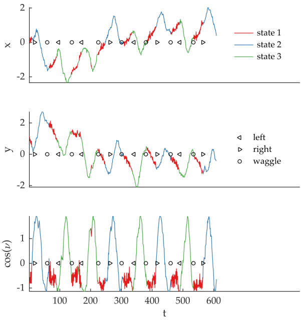
- *standardized* latent residuals

$$\tilde{\mathbf{z}}_t = (\mathbf{Q}^{(s_t^*)*})^{0.5} (\mathbf{z}_t^* - \mathbf{A}^{(s_t^*)*} \mathbf{z}_{t-1}^*)\ \forall\, t = 2, \ldots, n, \tag{11}$$

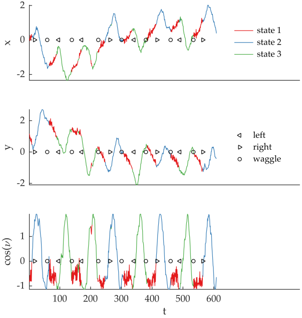
- *standardized* observation residuals (or innovations)

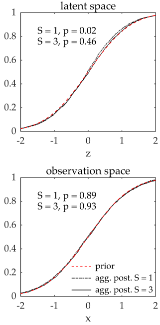
$$\tilde{\mathbf{x}}_t = (\mathbf{R}^*)^{0.5} (\mathbf{x}_t^{\text{obs}} - \mathbf{B}^* \mathbf{z}_t^*)\ \forall\, t = 2, \ldots, n. \tag{12}$$

- Four measurements of (x, y) coordinate and cosine and sine of head angle (ν)
- Three distinct dynamical regimes, namely, left turn, right turn and waggle



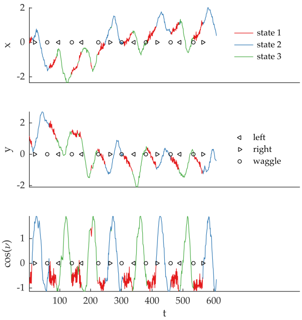
- Four measurements of (x, y) coordinate and cosine and sine of head angle (ν)
- Three distinct dynamical regimes, namely, left turn, right turn and waggle

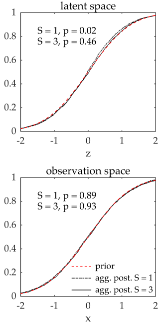


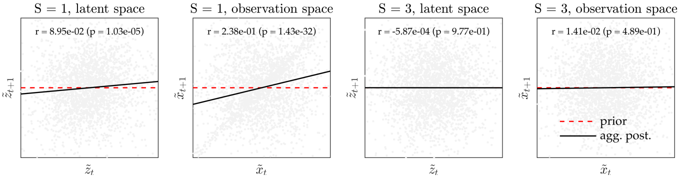


2. Honey Bee

- Four measurements of (x, y) coordinate and cosine and sine of head angle (ν)
- Three distinct dynamical regimes, namely, left turn, right turn and waggle



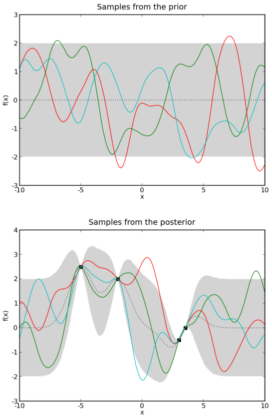




$$\vartheta, \zeta, \tau \sim p(\vartheta)\, p(\zeta)\, p(\tau), \tag{13}$$

$$f(x) \sim \mathcal{GP}(m(x \mid \vartheta), \kappa(x, x' \mid \zeta)), \tag{14}$$

$$y_i \sim \mathcal{N}(f(x_i), \tau^{-1}) \quad \forall i = 1, \ldots, n, \tag{15}$$



[SOURCE]

$$\vartheta, \zeta, \tau \sim p(\vartheta)\, p(\zeta)\, p(\tau), \tag{13}$$

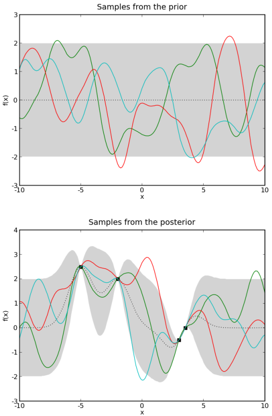
$$f(x) \sim \mathcal{GP}(m(x \mid \vartheta),\, \kappa(x, x' \mid \zeta)), \tag{14}$$

$$y_i \sim \mathcal{N}(f(x_i), \tau^{-1}) \quad \forall i = 1, \ldots, n, \tag{15}$$

$$\mathbf{m} = (m(\mathbf{x}_1 \mid \vartheta), \ldots, m(\mathbf{x}_n \mid \vartheta))^\top \tag{16}$$

$$\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j \mid \zeta) + \tau^{-1}\delta(\mathbf{x}_i, \mathbf{x}_j) \tag{17}$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \tag{18}$$



[SOURCE]

3. Gaussian Process Regression

$$\vartheta, \zeta, \tau \sim p(\vartheta)\, p(\zeta)\, p(\tau), \tag{13}$$

$$f(x) \sim \mathcal{GP}(m(x \mid \vartheta), \kappa(x, x' \mid \zeta)), \tag{14}$$

$$y_i \sim \mathcal{N}(f(x_i), \tau^{-1}) \quad \forall i = 1, \ldots, n, \tag{15}$$

$$\mathbf{m} = (m(\mathbf{x}_1 \mid \vartheta), \ldots, m(\mathbf{x}_n \mid \vartheta))^\top \tag{16}$$

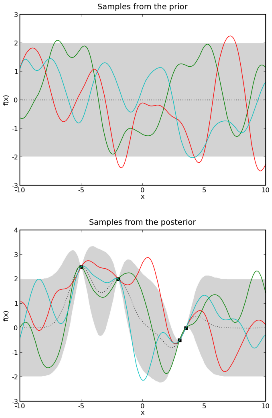
$$\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j \mid \zeta) + \tau^{-1}\delta(\mathbf{x}_i, \mathbf{x}_j) \tag{17}$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \tag{18}$$

$$\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top \tag{19}$$

$$\mathbf{c} = \mathbf{U}^\top(\mathbf{y} - \mathbf{m}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}) \tag{20}$$

$$\mathbf{z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\top(\mathbf{y} - \mathbf{m}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{21}$$



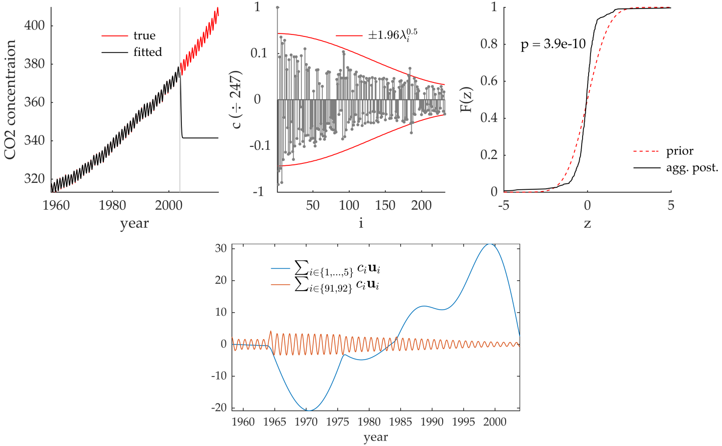
[SOURCE]

Given ν^*, ζ^*, τ^*

$$\{z_1^*, \ldots, z_n^*\} \quad \sim \quad \mathcal{N}(0, 1) \tag{22}$$

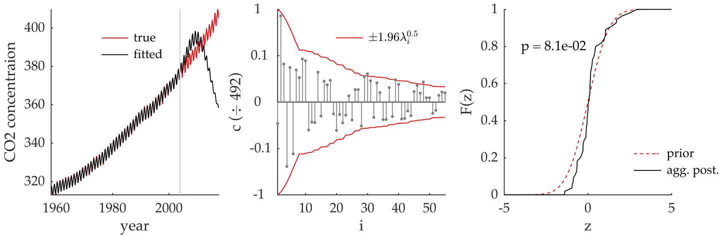
3. CO2 Emmision

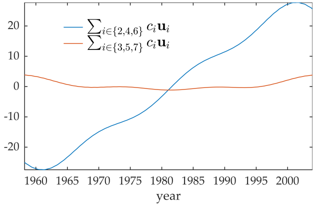
$$\kappa_{\mathrm{se}}(x, x' \mid \zeta) = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right) \tag{23}$$



3. CO2 Emmision

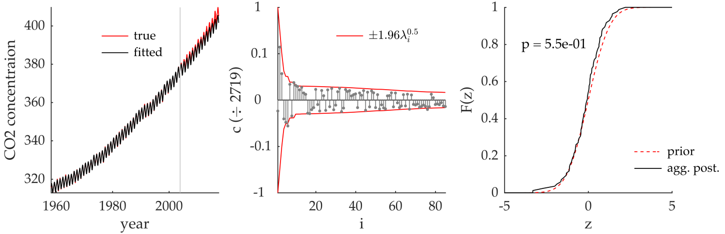
$$\kappa_{\text{pe}}(x, x' \mid \zeta) = \sigma_f^2 \exp\left(-\frac{2\sin^2(\pi(x-x')/p)}{l_p^2}\right)\exp\left(-\frac{(x-x')^2}{2l_d^2}\right) \tag{23}$$





3. CO2 Emmision

$$\kappa_{\text{se}}(x, x' \mid \zeta_s) + \kappa_{\text{se}}(x, x' \mid \zeta_l) + \kappa_{\text{pe}}(x, x' \mid \zeta) \tag{23}$$



Discussion

- Model criticism goodness-of-fit and graphical illustration for understanding a limitations of the model with the hope that a better model can be found
- The term *model criticism* is preferred over *model validation* and *model checking* since the former has a more active tone of looking to discover problems, while the latter may seem a more passive activity that does not expect to uncover any problems O'Hagan [2003, p423].
- 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].
- Aggregated Posterior Check complements Posterior Predictive Check by criticising the latent space rather than the observation space, and has been used in the literature in different forms Meulders et al. [1998], Buccigrossi and Simoncelli [1999], O'Hagan [2003], Tang et al. [2012]

Discussion

- Model criticism goodness-of-fit and graphical illustration for understanding a limitations of the model with the hope that a better model can be found
- The term *model criticism* is preferred over *model validation* and *model checking* since the former has a more active tone of looking to discover problems, while the latter may seem a more passive activity that does not expect to uncover any problems O'Hagan [2003, p423].
- 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].
- Aggregated Posterior Check complements Posterior Predictive Check by criticising the latent space rather than the observation space, and has been used in the literature in different forms Meulders et al. [1998], Buccigrossi and Simoncelli [1999], O'Hagan [2003], Tang et al. [2012]

Sohan Seth, Iain Murray, and Christopher K. I. Williams.
Model Criticism in Latent Space.
Bayesian Analysis, 14(3):703–725, 2019.
`https://projecteuclid.org/euclid.ba/1560240024`.

References

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

G. E.P Box. Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society*, 143(4):383–430, 1980.

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

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

V. E. Johnson. Bayesian model assessment using pivotal quantities. *Bayesian Anal.*, 2(4):719–733, 12 2007. doi: 10.1214/07-BA229.

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

A. O'Hagan. HSSS Model Criticism. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 422–444. Oxford University Press, 2003.

D. B. Rubin. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics*, 12:1151–1172, 1984.

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

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

Thank you for your patience!