Diagnostics for Conditional Density Models and Bayesian Inference Algorithms (Supplementary material)

David Zhao[1] **Niccolò Dalmasso**[1] **Rafael Izbicki**[2] **Ann B. Lee**[1]

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]Department of Statistics, Federal University of São Carlos (UFSCar), São Carlos, Brazil

SUPPLEMENTARY MATERIALS

A: PROOFS

In this section, we show proofs of the results stated in the paper.

Proof of Theorem 1. Let $z = g(\mathbf{x})$ and $Z = g(\mathbf{X})$. Notice Equation 3 implies $\widehat{F}(Y|\mathbf{x}) = F(Y|g(\mathbf{x})) = F(Y|z)$, and thus

$$\widehat{F}(Y|\mathbf{X}) = F(Y|g(\mathbf{X})) = F(Y|Z) \tag{1}$$

Thus, if $(\mathbf{X}, Y) \sim F_{\mathbf{X},Y}$ then, for every $0 \leq a \leq 1$,

$$\begin{aligned}
&\mathbb{P}(\text{PIT}(Y, \mathbf{X}) \leq a) = \mathbb{P}(\widehat{F}(Y|\mathbf{X}) \leq a) \\
&= \int_{\mathcal{Z}} \mathbb{P}(\widehat{F}(Y|\mathbf{X}) \leq a|Z = z) f(z) dz \\
&= \int_{\mathcal{Z}} \mathbb{P}(F(Y|Z) \leq a|Z = z) f(z) dz \qquad \text{(Eq. 1)} \\
&= \int_{\mathcal{Z}} \mathbb{P}(F(Y|z) \leq a|Z = z) f(z) dz \\
&= \int_{\mathcal{Z}} \mathbb{P}(Y \leq F^{-1}(a|z)|Z = z) f(z) dz \\
&= \int_{\mathcal{Z}} F(F^{-1}(a|z)|Z = z) f(z) dz = \int_{\mathcal{Z}} a f(z) dz = a.
\end{aligned}$$

□

Proof of Theorem 2. Assume that $\widehat{f}(y|\mathbf{x}) = f(y|\mathbf{x})$. It follows that, for any $0 < \alpha < 1$,

$$\begin{aligned}
\mathbb{P}\left(\text{PIT}(Y; \mathbf{X}) < \alpha|\mathbf{x}\right) &= \mathbb{P}\left(F_{Y|\mathbf{x}}(Y) \leq \alpha|\mathbf{x}\right) \\
&= \mathbb{P}\left(Y \leq F^{-1}_{Y|\mathbf{x}}(\alpha)|\mathbf{x}\right) \\
&= F_{Y|\mathbf{x}}\left(F^{-1}_{Y|\mathbf{x}}(\alpha)\right) \\
&= \alpha,
\end{aligned}$$

which shows that the distribution of $\text{PIT}(Y; \mathbf{X})$, conditional on $\mathbf{x}$, is uniform. Now, assume that $\mathbb{P}\left(\text{PIT}(Y; \mathbf{X}) < \alpha|\mathbf{x}\right) = \alpha$ for every $0 < \alpha < 1$ and let $\widehat{F}_{y|\mathbf{x}}(y) = \int_{-\infty}^{y} \widehat{f}(y'|\mathbf{x}) dy'$. Then

$$\begin{aligned}
\alpha &= \mathbb{P}\left(\text{PIT}(Y; \mathbf{X}) < \alpha|\mathbf{x}\right) \\
&= \mathbb{P}\left(\widehat{F}_{Y|\mathbf{x}}(Y) \leq \alpha|\mathbf{x}\right) \\
&= \mathbb{P}\left(Y \leq \widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha)|\mathbf{x}\right) \\
&= F_{Y|\mathbf{x}}\left(\widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha)\right).
\end{aligned}$$

It follows that $F_{Y|\mathbf{x}}\left(\widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha)\right) = \alpha$, and thus

$$\widehat{F}^{-1}_{Y|\mathbf{x}}(\alpha) = F^{-1}_{Y|\mathbf{x}}(\alpha) \ \forall \alpha \in (0, 1).$$

The conclusion follows from the fact that the CDF characterizes the distribution of a random variable. □

Proof of Corollary 1. Notice that $r_\alpha(\mathbf{x}) = \mathbb{E}\left[Z^\alpha|\mathbf{x}\right] = \mathbb{P}\left(\text{PIT}(Y; \mathbf{X}) < \alpha|\mathbf{x}\right)$. It follows that $r_\alpha(\mathbf{x}) = \alpha$ for every $\alpha \in (0, 1)$ if, and only if, the distribution of $\text{PIT}(\mathbf{Y}; \mathbf{X})$, conditional on $\mathbf{X}$, is uniform over $(0, 1)$. The conclusion follows from Theorem 2. □

Theorem 4 (HPD values are insensitive to covariate transformations). *Let $(\mathbf{X}, \mathbf{Y}) \sim F_{\mathbf{X},\mathbf{Y}}$. If there exists a function $g : \mathcal{X} \to \mathcal{Z}$ such that $\widehat{f}(\mathbf{y}|\mathbf{x}) = f(\mathbf{y}|g(\mathbf{x}))$, then HPD$(\mathbf{Y}; \mathbf{X}) \sim Unif(0, 1)$.*

Proof of Theorem 4. Under the assumption we can rewrite the HPD value as:

$$\begin{aligned}
\text{HPD}(\mathbf{y}, \mathbf{x}) &= \int_{\mathbf{y}': f(\mathbf{y}'|g(\mathbf{x})) > f(\mathbf{y}|g(\mathbf{x}))} f(\mathbf{y}'|g(\mathbf{x})) dy' \\
&= \int_{y': f(\mathbf{y}'|\mathbf{z}) > f(\mathbf{y}|\mathbf{z})} f(\mathbf{y}'|\mathbf{z}) dy' = \text{HPD}(\mathbf{y}, \mathbf{z}),
\end{aligned}$$

with $g(\mathbf{x}) = \mathbf{z}$. Following the proof structure by Harrison et al. [2015] closely, we define the random variable $\xi_{\mathbf{z},\mathbf{y}} = \text{HPD}(\mathbf{z}, \mathbf{y})$, equipped with the probability density function $h : (\mathcal{Z} \times \mathcal{Y}) \to \mathbb{R}$. Dropping the subscripts for simplicity, let $\xi^* = \text{HPD}(\mathbf{z}^*, \mathbf{y}^*)$ the HPD value of a specific pair $(\mathbf{z}^*, \mathbf{y}^*)$; ξ^* is the probability mass of f above the level set $f(\mathbf{y}^*|\mathbf{z}^* = g(\mathbf{x}^*))$. Without loss of generality, if we show that $h(\xi^*) = 1$ we can conclude that $\xi(y, z)$ is uniformly distributed $U[0, 1]$. Using the fundamental theorem of calculus we can write:

Algorithm 3 P-values for Global Coverage Test

Require: conditional density model $\widehat{f}$; test data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$; regression estimator $\widehat{r}$; number of null training samples B

Ensure: estimated p-value $\widehat{p}(\mathbf{x})$ across all $\mathbf{x} \in \mathcal{X}$

```
// Compute test statistic over $\mathbf{X}_1, \dots, \mathbf{X}_n$:
Compute values $\text{PIT}(Y_1; \mathbf{X}_1), \dots, \text{PIT}(Y_n; \mathbf{X}_n)$
$G \leftarrow$ grid of $\alpha$ values in $(0, 1)$.
for $\alpha$ in $G$ do
    Compute indicators $Z_1^\alpha, \dots, Z_n^\alpha$
    Train regression method $\widehat{r}_\alpha$ on $\{\mathbf{X}_i, Z_i^\alpha\}_{i=1}^n$
end for
Compute test statistic $S = \frac{1}{n}\sum_{i=1}^n T(\mathbf{X}_i)$
// Recompute test statistic under null distribution:
for $b$ in $1, \dots, B$ do
    Draw $U_1^{(b)}, \dots, U_n^{(b)} \sim \text{Unif}[0, 1]$.
    for $\alpha$ in $G$ do
        Compute indicators $\{Z_{\alpha,i}^{(b)} = \mathbb{I}(U_i^{(b)} < \alpha)\}_{i=1}^n$
        Train regression method $\widehat{r}_\alpha^{(b)}$ on $\{\mathbf{X}_i, Z_{\alpha,i}^{(b)}\}_{i=1}^n$
    end for
    Compute $T^{(b)}(\mathbf{X}_i) := \frac{1}{|G|}\sum_{\alpha \in G}(\widehat{r}_\alpha^{(b)}(\mathbf{X}_i) - \alpha)^2$ for $i = 1, \dots, n$
    Compute $S^{(b)} := \frac{1}{n}\sum_{i=1}^n T^{(b)}(\mathbf{X}_i)$
end for
return $\widehat{p}(\mathbf{x}) := \frac{1}{B}\sum_{b=1}^B \mathbb{I}\left(S < S^{(b)}\right)$
```

$$
\begin{aligned}
h(\xi^*) &= \frac{\partial}{\partial \xi^*} \int_{-\infty}^{\xi^*} g(\epsilon) d\epsilon \\
&= \frac{\partial}{\partial \xi^*} \int_{-\infty}^{\xi^*} \int_{\mathcal{Z} \times \mathcal{Y}} \delta(\xi(y, z) - \epsilon) dF(z, y) d\epsilon \\
&= \frac{\partial}{\partial \xi^*} \int_{\mathcal{Z} \times \mathcal{Y}} \Phi(\xi(y, z) - \xi^*) dF(z, y) \\
&= \frac{\partial}{\partial \xi^*} \int_{\mathcal{Z}} \left[\int_{\mathcal{Y}} \Phi(\xi(y, z) - \xi^*) f(y|z) dy \right] f(z) dz \\
&= \frac{\partial}{\partial \xi^*} \int_{\mathcal{Z}} \xi^* f(z) dz = \frac{\partial}{\partial \xi^*} \xi^* = 1
\end{aligned}
$$

where Φ is the Heavyside function, which is 1 when the argument is positive and 0 otherwise. □

Proof of Theorem 3. Under the null hypothesis $H_0(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$ we have that:

$$
\begin{aligned}
\text{HPD}(\mathbf{y}; \mathbf{x}) &= \int_{\mathbf{y}' : \widehat{f}(\mathbf{y}'|\mathbf{x}) \geq \widehat{f}(\mathbf{y}|\mathbf{x})} \widehat{f}(\mathbf{y}'|\mathbf{x}) d\mathbf{y} \qquad (2) \\
&= \int_{\mathbf{y}' : f(\mathbf{y}'|\mathbf{x}) \geq f(\mathbf{y}|\mathbf{x})} f(\mathbf{y}'|\mathbf{x}) d\mathbf{y}. \qquad (3)
\end{aligned}
$$

Applying the results about uniformity of HPD for $f(\cdot|\mathbf{x})$ from Harrison et al. [2015, Section A.2] (also reproduced in the proof of Theorem 4) proves the theorem.

□

B: GLOBAL COVERAGE TEST

Algorithm 3 describes our procedure for testing global consistency (see Definition 1 in the paper) using a Monte Carlo sampling strategy.

C: EXAMPLE 1: OMITTED VARIABLE BIAS IN CDE MODELS

In this section we show the results of the local test on Example 1 for model $\widehat{f}_2$, which passes the global test.

Figure 7, right panel, shows p-values from LCTs across the feature space for the model $\widehat{f}_2$. Unlike model $\widehat{f}_1$, which was fit on X_1 alone, $\widehat{f}_2$ was fit on both X_1 and X_2. Hence, $\widehat{f}_2$ is able to pass all tests, with local P-P plots indicating a good fit (with two examples shown in the Figure 7, left panel).

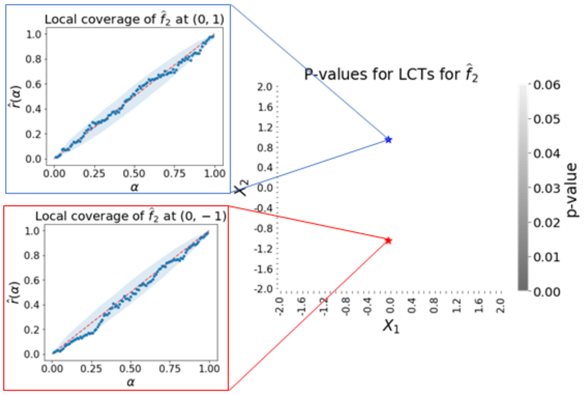

Figure 7: P-values for LCTs for $\widehat{f}_2$ in Example 1 suggest an adequate fit everywhere in the feature space; local coverage plots at selected points also suggest a good fit.

K	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
KL loss	-0.729	-0.885	-0.915	-0.906	-0.897	-0.917	-0.906	-0.911	-0.905

Table 1: The KL divergence loss indicates that the number of mixture components in the ConvMDN approximation of the posterior in Example 2 should be $K = 7$.

D: EXAMPLE 2: CONDITIONAL NEURAL DENSITY MODELING FOR GALAXY IMAGES

Figure 8 shows the true conditional densities of the simulated "redshift" Z vs. the axis ratio λ of the corresponding galaxy image.

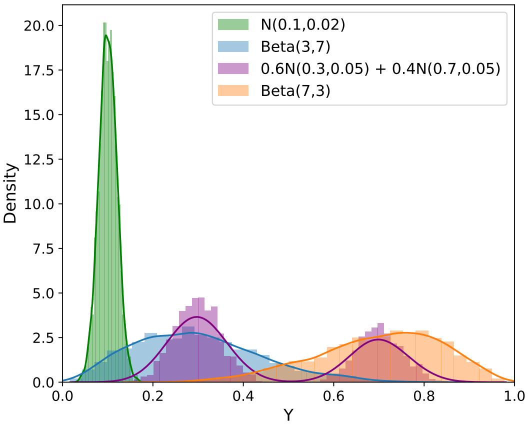

Figure 8: We assign a unimodal distribution of "redshift" Z for to the galaxy population with $\lambda = 0.8$, and higher, more skewed and bimodal distributions of Z to the populations with $\lambda = 0.7, 0.6, 0.5$.

E: EXAMPLE 3: POSTERIOR INFERENCE FOR GALAXY IMAGES

Table 1 reports the KL divergence loss over a test set of 1000 galaxy images for a ConvMDN model with K components, for $K = 2, ..., 10$. The KL loss indicates that $K = 7$ is the optimal choice. However, in the paper we show that this model fails to pass our GCT and therefore is not a good approximation of the true conditional density. Figure 6 in the paper also shows how to use our LCTs and P-P plots to diagnose the inadequacies in the fit.

F: EXAMPLE 4: CONDITIONAL DENSITY MODELS WITH MULTIVARIATE RESPONSE

For multivariate response $\mathbf{Y}$, we can assess the quality of fit of $\widehat{f}$ through highest predictive density (HPD) values, as described in Section 3.3. Our method still yields interpretable diagnostics, but the interpretation of HPD values differs from that of PIT values. If a local P-P plot shows estimated HPD values $\widehat{r}_\alpha$ that are too high relative to α, this suggests that the model is overdispersed relative to the true density. HPD values that are too low could suggest an underdispersed model, or be a symptom of model misspecification: if the estimated density is systematically biased (i.e. not centered at the same location as the true density), the observed values Y will disproportionately represent lower density contours of the true density.

In this example, we draw $\mathbf{X} = (X_1, X_2) \sim \text{Unif}[0,1]^2$, and then define a bivariate response $\mathbf{Y} = (Y_1, Y_2)$ as follows:

$$\mathbf{Y}|\mathbf{X} \sim \begin{cases} N((X_1, X_2), I_2), & X_2 \in [1,2] \\ N((X_1, X_2), 0.25 I_2), & X_2 \in [0,1] \\ t_4 \text{ centered at } (X_1, X_2), & X_2 \in [-1,0] \\ t_4 \text{ centered at } (X_1+1, X_2+1), & X_2 \in [-2,-1] \end{cases}$$

where I_2 is the identity matrix. See Figure 9 for an illustration of how the true conditional density $f(\mathbf{y}|\mathbf{x})$ varies across the feature space. For illustration, we choose the model $\widehat{f}(\cdot|\mathbf{x}) = N((x_1, x_2), 1)$ in all four regions. This model perfectly fits the true density when $x_2 \in [1,2]$, and is misspecified in the other cases. We evaluate HPD values at 1000 test points to run our diagnostic framework.

Figure 10 summarizes the results of our diagnostics. First, we perform the GCT, which rejects the global null with $p < 0.001$. We then perform LCTs across the feature space for $\mathbf{X}$; the resulting p-values are shown in the center panel. As expected, LCTs indicate a good fit when $\widehat{f}$ is correct, and a poor fit in most regions where $\widehat{f}$ is misspecified. Investigating further with local P-P plots enables us to detect overcoverage and undercoverage of HPD regions at specific locations in the feature space. Overcoverage of the true $\mathbf{Y}$ by the HPD region means the α-HPD set for $\widehat{f}$ is too large, so observed HPD values are too low: this indicates that $\widehat{f}$ is overdispersed locally (as in the top right example). Conversely, undercoverage by the HPD region means the α-HPD set for $\widehat{f}$ does not cover enough of the true density mass of f, so observed HPD values are too high: this can be caused by $\widehat{f}$ being underdispersed or biased locally (as in the bottom right example).

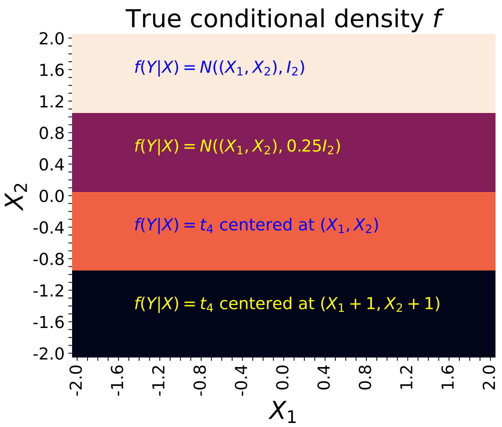

Figure 9: The true conditional density $f(\mathbf{y}|\mathbf{x})$ has different forms in four different regions of the feature space, whereas we assume the same model $\widehat{f}(\mathbf{y}|\mathbf{x}) = N((x_1, x_2), 1)$ across feature space. When $X_2 \in [1, 2]$, the model $\widehat{f}$ is correctly specified. When $X_2 \in [0, 1]$, $\widehat{f}$ is overdispersed relative to the true density f. When $X_2 \in [-1, 0]$, $\widehat{f}$ is slightly underdispersed relative to the true density f. When $X_2 \in [-2, -1]$, $\widehat{f}$ is both biased and slightly underdispersed relative to the true density f.

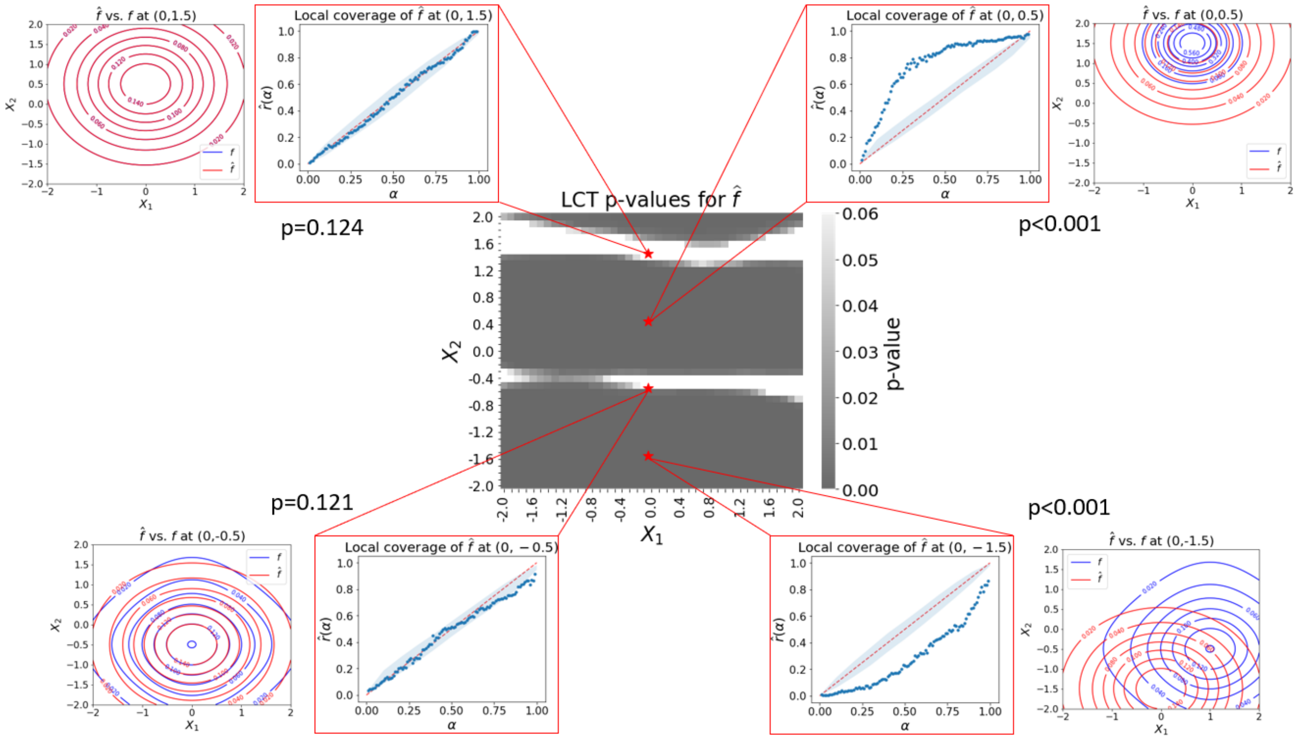

Figure 10: New diagnostics for Example 4. P-values for LCTs for $\widehat{f}$ indicate a poor fit for values of X where $X_2 \in [0, 1]$ or $X_2 \in [-2, -1]$ (see center panel). Amortized local P-P plots at selected points show the HPD level sets of $\widehat{f}$ as overdispersed for $X_2 \in [0, 1]$, and underdispersed or biased for $X_2 \in [-2, -1]$. In contrast, the HPD level sets are well estimated at significance level $\alpha = 0.05$ for $X_2 \in [1, 2]$ and $X_2 \in [-1, 0]$. (Gray regions represent 95% confidence bands under the null.) Contour plots show the model $\widehat{f}$ vs. the true (unknown) conditional density f at the selected points. $\widehat{f}$ is clearly overdispersed at $(0, 0.5)$ and systematically biased at $(0, -1.5)$. The model perfectly fits the density at $(0, 1.5)$, and has barely detectable underdispersion at $(0, -0.5)$. (*Note:* The contour plots requires knowledge of the true f, which would not be available to the practitioner.)

References

Diana Harrison, David Sutton, Pedro Carvalho, and Michael Hobson. Validation of Bayesian posterior distributions using a multidimensional Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society*, 451(3):2610–2624, 06 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv1110.