Approximate Bayesian Computation with Statistical Distances for Model Selection

Clara Grazian

School of Mathematics and Statistics, University of Sydney

clara.grazian@sydney.edu.au

OWABI Seminar Series

Online

27 November 2025

The problem of model choice

The selection of a model that best explains a dataset or process is central to statistics and relevant across disciplines.

While no model can fully capture all nuances in complex phenomena, approximating the true data-generating process (DGP) can yield valuable insights [Molnar et al., 2022, Gelman and Shalizi, 2013].

However, many models involve complex interdependencies or latent variables that render their likelihood functions intractable [Martin et al., 2024] $\rightarrow$ ABC.

ABC for inference

1. To mitigate the curse of dimensionality, ABC often relies on low-dimensional summary statistics but this may be particularly problematic for model selection [Robert et al., 2011].
2. Forbes et al. [2022] introduced a framework that constructs surrogate posteriors using finite Gaussian mixtures
3. other ABC methods compare entire empirical distributions using statistical distances [Park et al., 2016, Bernton et al., 2019, Nguyen et al., 2020, Frazier, 2020, Drovandi and Frazier, 2022]
4. machine learning can be used to learn informative low-dimensional summaries automatically [Blum and François, 2010, Sheehan and Song, 2016, Jiang et al., 2017]

Background

Data: Let $\boldsymbol{y} = (y_1, ..., y_n)^T \in \mathscr{Y}^n \subseteq \mathbb{R}^n$ denote the observed data, generated from $P^{(n)}_{\boldsymbol{\theta}_0}$ within a model family $\mathscr{P} = \{P^{(n)}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$.

Prior: Model M is parameterized by $\boldsymbol{\theta}$ with prior distribution $\pi(\boldsymbol{\theta})$

Likelihood: The likelihood function is $p(\boldsymbol{y}|\boldsymbol{\theta})$

Posterior: The posterior distribution is $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

Marginal: The marginal likelihood is as $p(\boldsymbol{y}) = \int_\Theta p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$.

Model choice in a Bayesian setting

- a set of K candidate models $M_1, \ldots, M_K$
- each model M_k has parameters $\boldsymbol{\theta}_k \in \Theta_k \subseteq \mathbb{R}^{d_k}$
- prior distribution for parameters $\pi_k(\boldsymbol{\theta}_k)$ and prior probabilities for models $\pi(M_k) = \pi(M = M_k)$ such that $\sum_{k=1}^{K} \pi(M_k) = 1$
- likelihood function $p_k(\boldsymbol{y}|\boldsymbol{\theta}_k) = p(\boldsymbol{y}|\boldsymbol{\theta}_k, M_k)$
- the posterior probability for model M_k is:

$$\pi(M_k|\boldsymbol{y}) = \frac{p_k(\boldsymbol{y})\pi(M_k)}{\sum_{j=1}^{K} p_j(\boldsymbol{y})\pi(M_j)} \quad \forall\, k = 1, ..., K,$$

The true DGP $P^{(n)}_{M_0,\boldsymbol{\theta}_0}$ is assumed to belong to the broader set $\mathscr{P} = \{P^{(n)}_{M_k,\boldsymbol{\theta}_k} \mid \boldsymbol{\theta}_k \in \Theta_k \subseteq \mathbb{R}^{d_k}, k \in \{1, \ldots, K\}\}$.

ABC for Model Choice

Model choice via ABC introduces challenges beyond standard inference [Robert et al., 2011]. Consider the Bayes factor between Model M_1 and M_2; the ABC approximation to B_{12} is

$$\hat{B}_{12}(\mathbf{y}) = \frac{\pi(M = M_2)\sum_{t=1}^{T} \mathbb{I}_{m^{(t)}=M_1} \mathbb{I}_{\rho\{\eta(\mathbf{z}^{(t)}),\mathbf{y})\leq\varepsilon\}}}{\pi(M = M_1)\sum_{t=1}^{T} \mathbb{I}_{m^{(t)}=M_1} \mathbb{I}_{\rho\{\eta(\mathbf{z}^{(t)}),\mathbf{y})\leq\varepsilon\}}}$$

By letting $T \to \infty$ and $\varepsilon \to 0$, we have

$$B_{12}^{\eta}(\mathbf{y}) = \frac{\int \pi_1(\theta_1) f_1^{\eta}(\eta(\mathbf{y}|\theta_1)) d\theta_1}{\int \pi_2(\theta_2) f_2^{\eta}(\eta(\mathbf{y}|\theta_2)) d\theta_2}$$

But this means that there is a mismatch between the true Bayes factor and the ABC Bayes factor

$$B_{12}(\mathbf{y}) = \frac{g_1(\mathbf{y}) \int \pi_1(\theta_1) f_1^{\eta}(\eta(\mathbf{y}|\theta_1)) d\theta_1}{g_2(\mathbf{y}) \int \pi_2(\theta_2) f_2^{\eta}(\eta(\mathbf{y}|\theta_2)) d\theta_2} = \frac{g_1(\mathbf{y})}{g_2(\mathbf{y})} B_{12}^{\eta}(\mathbf{y}).$$

Alternatives: Marin et al. [2013], Barnes et al. [2012], Prangle et al. [2014].

Discrepancy-based ABC-MC

To incorporate statistical distances into model selection, the ABC distance metric $\rho(\cdot,\cdot)$ is replaced with a discrepancy measure $\mathscr{D}(\cdot,\cdot)$ that operates on the probability measures of the observed data $\boldsymbol{y}$ and simulated data $\boldsymbol{z}$.

Denote $\mu_{0,\boldsymbol{\theta}_0}$ as the measure associated with the true distribution $P^{(n)}_{M_0,\boldsymbol{\theta}_0}$, and $\mu_{k^*,\boldsymbol{\theta}^*_{k^*}}$ as the measure associated with the simulated data's distribution $P^{(n)}_{M_{k^*},\boldsymbol{\theta}_{k^*}} \in \mathscr{P}$.

The empirical distributions of $\boldsymbol{y}$ and $\boldsymbol{z}$ replace the true measures, and are similarly defined as $\hat{\mu}_{0,\boldsymbol{\theta}_0} = n^{-1}\sum_{i=1}^{n}\delta_{y_i}$ and $\hat{\mu}_{k^*,\boldsymbol{\theta}^*_{k^*}} = n^{-1}\sum_{i=1}^{n}\delta_{z_i}$ respectively.

Discrepancy-based ABC-MC

Algorithm 1 Discrepancy-based ABC-MC

Input: Given a set of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, a set of possible models $\{M_1, M_2, \ldots, M_K\}$, a tolerance level ε, and a discrepancy metric $\mathscr{D}(\cdot, \cdot)$:
for $i = 1, \ldots, N$ **do**
    repeat
        Generate M_{k^*} from $\pi(M_k)$, $k = 1, \ldots, K$.
        Generate $\boldsymbol{\theta}_{k^*}$ from $\pi_{k^*}(\boldsymbol{\theta}_{k^*})$.
        Generate $\boldsymbol{z} = (z_1, ..., z_n)^T$ from $P^{(n)}_{M_{k^*}, \boldsymbol{\theta}_{k^*}}$.
    until $\mathscr{D}(\hat{\mu}_{0,\boldsymbol{\theta}_0}, \hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}) \leq \varepsilon$,
    Set $M^{(i)} = M_{k^*}$ and $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_{k^*}$.
end for
Output: A set of values $(M^{(1)}, \ldots, M^{(N)})$ and $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)})$ from $\pi_\varepsilon(M_k|\boldsymbol{y})$ and $\pi_{\varepsilon,k}(\boldsymbol{\theta}_k|\boldsymbol{y}, M_k)$, for $k = 1, \ldots, K$, respectively.

What distances?

The maximum mean discrepancy (MMD) [Park et al., 2016] compares embedded empirical distributions in a reproducing kernel Hilbert space, using a kernel $g : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$,

$$MMD^2(\mu_{0,\boldsymbol{\theta}_0}, \mu_{k^*,\boldsymbol{\theta}_{k^*}}) = \mathbb{E}[g(y_1, y_2)] + \mathbb{E}[g(z_1, z_2)] - 2\mathbb{E}[g(y_1, z_1)].$$

The unbiased empirical estimate is:

$$MMD^2(\hat{\mu}_{0,\boldsymbol{\theta}_0}, \mu_{k^*,\boldsymbol{\theta}_{k^*}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} g(y_i, y_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} g(z_i, z_j) + - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g(y_i, z_j).$$

What distances?

Bernton et al. [2019] proposed using the Wasserstein distance in ABC.

For univariate data and assuming that $P^{(n)}_{M_0,\boldsymbol{\theta}_0}$ and $P^{(n)}_{M_{k^*},\boldsymbol{\theta}_{k^*}}$ have finite p-th moment with $p \geq 1$, the p-Wasserstein distance is

$$\mathscr{W}_p(\mu_{0,\boldsymbol{\theta}_0}, \mu_{k^*,\boldsymbol{\theta}_{k^*}}) = \left(\int_0^1 |F^{-1}_{\mu_{0,\boldsymbol{\theta}_0}}(\lambda) - F^{-1}_{\mu_{k^*,\boldsymbol{\theta}_{k^*}}}(\lambda)|^p d\lambda \right)^{1/p},$$

where $F_{\mu_{0,\boldsymbol{\theta}_0}}(\cdot)$ and $F_{\mu_{k^*,\boldsymbol{\theta}_{k^*}}}(\cdot)$ are the cumulative distribution.

The empirical p-Wasserstein distance, with $p = 1$, is computed by comparing order statistics:

$$\mathscr{W}_1(\hat{\mu}_{0,\boldsymbol{\theta}_0}, \hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}) = n^{-1} \sum_{i=1}^{n} |y_{(i)} - z_{(i)}|.$$

What distances?

The Cramér–von Mises (CvM) distance [Frazier, 2020] measures the L_2 difference between the empirical CDF $\hat{F}_{\mu_{k^*},\boldsymbol{\theta}_{k^*}}$ and the CDF of the theoretical distribution $F_{\mu_0,\boldsymbol{\theta}_0}$

$$\mathscr{C}^2(\mu_{0,\boldsymbol{\theta}_0},\hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}) = \int_{\mathscr{Y}} \left[\hat{F}_{\mu_{k^*},\boldsymbol{\theta}_{k^*}}(y) - F_{\mu_0,\boldsymbol{\theta}_0}(y)\right] dF_{\mu_0,\boldsymbol{\theta}_0}(y).$$

It is estimated by:

$$\mathscr{C}^2(\hat{\mu}_{0,\boldsymbol{\theta}_0},\hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}) = \frac{U}{2n^2} - \frac{4n^2-1}{12n},$$

where U involves the ranks of $\boldsymbol{y}$ and $\boldsymbol{z}$ in their pooled sample.

Theoretical justification: Consistency

Consider Algorithm 1 using

- the Wasserstein distance [Bernton et al., 2019],
- the candidate models are well-separated in the Wasserstein sense, and
- a sequence $\varepsilon_n \to 0$ (since with growing sample size the empirical distributions concentrate around their population laws),
- observations are i.i.d.

The ABC posterior probability of model $M_{k^*} \in \{M_1, \ldots, M_K\} = \mathscr{M}$ is

$$\pi_{\varepsilon_n}(M_{k^*} \mid \boldsymbol{y}) \propto \pi(M_{k^*}) \int_{\Theta_{k^*}} \pi(\boldsymbol{\theta}_{k^*} \mid M_{k^*}) \mathbb{P}\left\{ \mathscr{W}_p(\hat{\mu}_{0,\boldsymbol{\theta}_0}, \hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}) \leq \varepsilon_n \right\} d\boldsymbol{\theta}_{k^*}.$$

Under regularity conditions, ABC-Wass yields consistent parameter estimation [Bernton et al., 2019]. This property extends to model selection.

Theoretical justification: Consistency

Theorem (Model Selection Consistency of ABC-Wass)

Assume:

(A1) *(Identifiability): For all* $M_{k^*} \neq M_0$, *there exists* $\delta_{k^*} > 0$ *such that* $\inf_{\boldsymbol{\theta}_{k^*} \in \Theta_{k^*}} \mathscr{W}_p(P^{(n)}_{M_0,\boldsymbol{\theta}_0}, P^{(n)}_{M_{k^*},\boldsymbol{\theta}_{k^*}}) > \delta_{k^*}$.

(A2) *(Prior Positivity):* $\pi(M_0) > 0$ *and* $\pi_{M_0}(\boldsymbol{\theta}_0) > 0$.

(A3) *(Tail Behavior): For any sequence* $\varepsilon_n \to 0$ *with* $\varepsilon_n < \delta_{k^*}/2$ *for large* n

$$\int_{\Theta_{k^*} \setminus \mathscr{B}_{\delta_{k^*}}} \pi_{M_{k^*}}(\boldsymbol{\theta}_{k^*}) \cdot \mathbb{P}(\mathscr{W}_p(\hat{\mu}_{0,\boldsymbol{\theta}_0}, \hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}) \leq \varepsilon_n) d\boldsymbol{\theta}_{k^*} \to 0 \quad \text{as } n \to \infty.$$

(A4) *(Empirical Convergence):* $\mathscr{W}_p(\hat{\mu}_{0,\boldsymbol{\theta}_0}, P^{(n)}_{M_0,\boldsymbol{\theta}_0}) \to 0$ *and* $\mathscr{W}_p(\hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}, P^{(n)}_{M_{k^*},\boldsymbol{\theta}_{k^*}}) \to 0$ *almost surely as* $n \to \infty$, *and* $\Delta_n := \sup_{M_{k^*},\boldsymbol{\theta}_{k^*}} \mathbb{E}[\mathscr{W}_p(\hat{\mu}_{k^*,\boldsymbol{\theta}_{k^*}}, P^{(n)}_{M_k^*,\boldsymbol{\theta}_{k^*}})]$ *satisfies* $\Delta_n = o(\varepsilon_n)$.

Then, for any sequence $\varepsilon_n \to 0$ *such that* $\varepsilon_n < \min_{k^*} \delta_{k^*}/2$,

$$\pi_{\varepsilon_n}(M_0 \mid \boldsymbol{y}) \xrightarrow{p} 1 \quad \text{as } n \to \infty.$$

Theoretical justification: Consistency

- **Assumption (A1)** ensures that for any incorrect model $M_{k^*} \neq M_0$ and any parameter $\boldsymbol{\theta}_{k^*} \in \Theta_{k^*}$, the Wasserstein distance between the simulated data distribution $P^{(n)}_{M_{k^*},\boldsymbol{\theta}_{k^*}}$ and the true distribution $P^{(n)}_{M_0,\boldsymbol{\theta}_0}$ is bounded below by some fixed $\delta_{k^*} > 0$.
- **Assumption (A2)** requires that both the model prior and the parameter prior assign positive mass to a nighborhoods of the true model and true parameter.
- **Assumption (A3)** ensures that parameter values far (in Wasserstein distance) to $P^{(n)}_{M_0,\boldsymbol{\theta}_0}$ cannot produce empirical measures within the small threshold with non-negligible prior mass.
- **Assumption (A4)** holds for i.i.d. observations and in cases with a finite p-th moment.

Theoretical justification: Robustness

Theorem (Robustness to Model Misspecification)

Let the true data distribution $P^{(n)}_{M_0,\boldsymbol{\theta}_0}$ *be misspecified with respect to the candidate set*

$$\{P^{(n)}_{M_k,\boldsymbol{\theta}_k} : M_k \in \mathscr{M}, \boldsymbol{\theta}_k \in \Theta_k\}.$$

Suppose Assumptions (A1)-(A4) hold, and let $\varepsilon_n \to 0$, *at a rate slower than the stochastic convergence of the Wasserstein distance* $\mathscr{W}_p(\hat{\mu}_0, P^{(n)}_{M_0,\boldsymbol{\theta}_0})$. *Then the ABC-Wass posterior asymptotically concentrates on the model–parameter pair*

$$(M^\dagger, \boldsymbol{\theta}^\dagger) = \arg \min_{M_k \in \mathscr{M}, \boldsymbol{\theta}_k \in \Theta_k} \mathscr{W}_p(P^{(n)}_{M_0,\boldsymbol{\theta}_0}, P^{(n)}_{M_k,\boldsymbol{\theta}_k}).$$

Simulation study

- 3 ABC with distance metrics (ABC-CvM, ABC-MMD, and ABC-Wass)
- ABC-Stat
- NN, performing multi-task learning: classification (model selection) with categorical cross-entropy loss and regression (parameter estimation) using MSE loss. Loss weights are 1.0 (classification) and 0.5 (regression), prioritizing model identification. The architecture comprises five fully connected layers.
- ABC-QDA Prangle et al. [2014]
- ABC-SA [Gutmann et al., 2018]

Each experiment is repeated 100 times for sample sizes $n = 100$. For each observed dataset, 10^6 simulations.

Model priors: $\pi(M = M_k) = 1/K$ for $k = 1, \ldots, K$.

The threshold ε is chosen as the q-th percentile of the simulated distances [Biau et al., 2015].

Normal example

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be i.i.d. with $y_i \sim N(\theta, \sigma^2)$, assuming $\sigma^2 = 1$. The competing models correspond to $H_0 : \theta = \tilde{\theta}$ versus $H_1 : \theta \neq \tilde{\theta}$.

Synthetic datasets are generated from the true model $\mathcal{N}(\theta_0, 1)$ with $\theta_0 \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

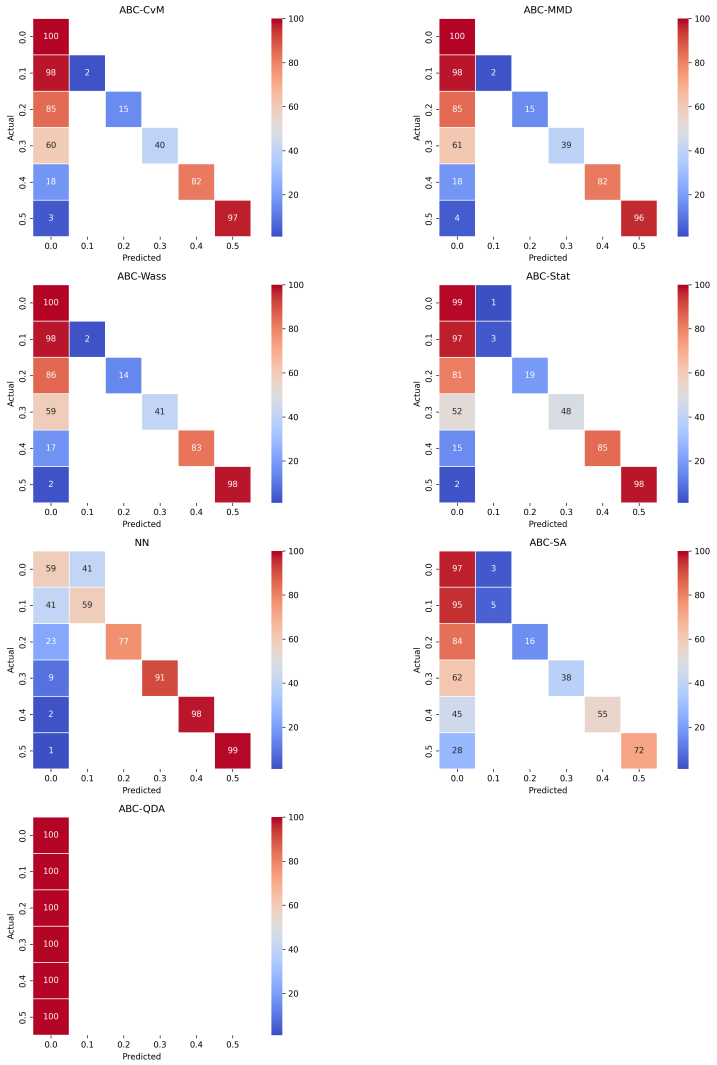
Under H_1, the prior is $\theta \sim N(\tilde{\theta}, 100)$, providing a weakly informative prior; under H_0, θ is fixed at $\tilde{\theta} = 0$.

For ABC-Stat, we use the sufficient statistic $\eta(\boldsymbol{y}) = \bar{y}$ and the Euclidean distance.

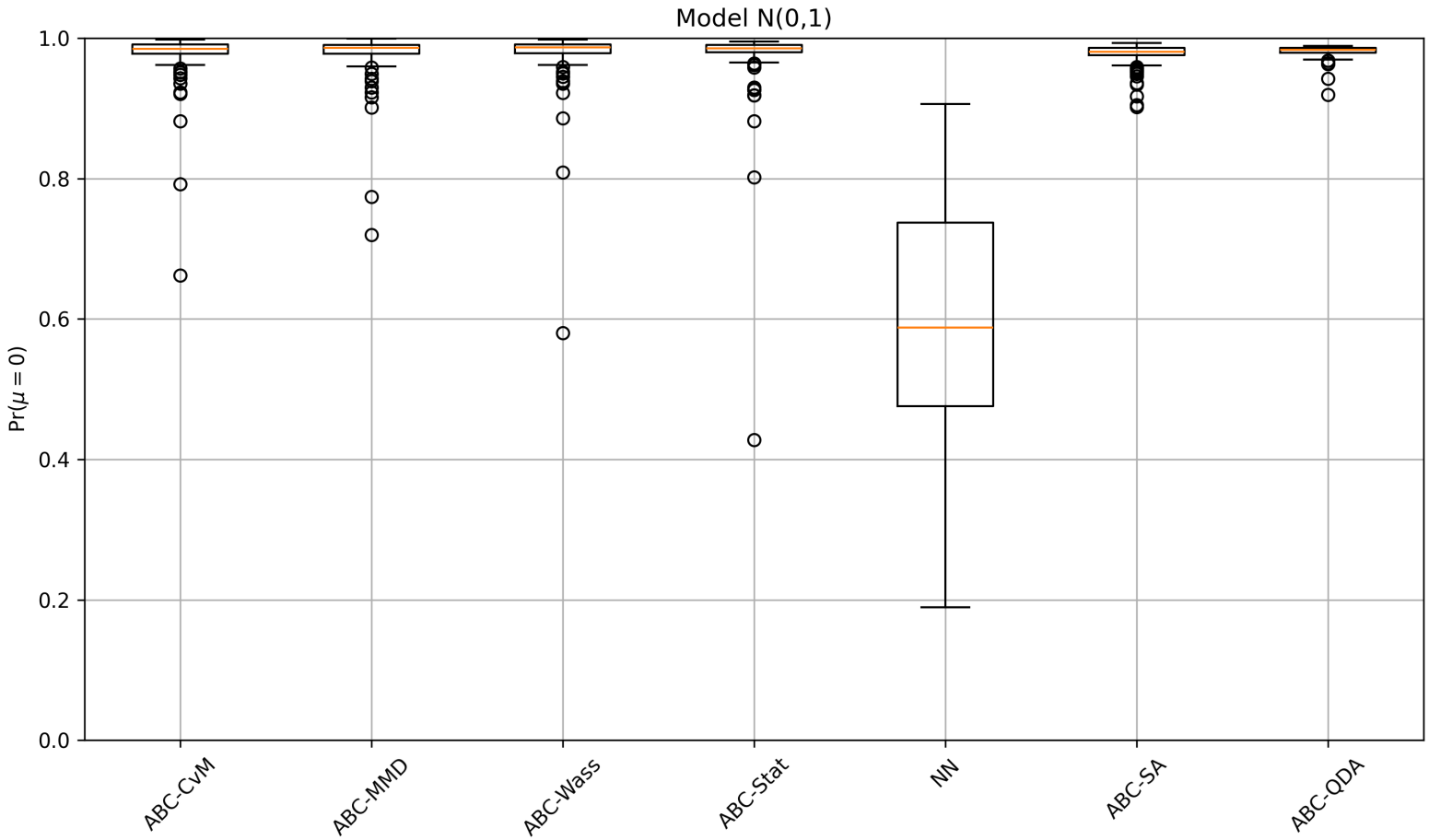
Normal example

		ABC						
		CvM	**MMD**	**Wass**	**Stat**	**NN**	**SA**	**QDA**
$\theta_0 = 0.0$	$\Pr(\theta = 0)$	0.975	0.976	0.975	0.973	0.599	0.976	0.981
	$\hat{\theta}$	0.000	0.000	0.000	0.002	0.042	0.000	0.000
	MSE	0.000	0.000	0.000	0.001	0.002	0.000	0.000
$\theta_0 = 0.1$	$\Pr(\theta = 0)$	0.931	0.933	0.929	0.412	0.925	0.948	0.974
	$\hat{\theta}$	0.007	0.006	0.007	0.010	0.052	0.000	0.000
	MSE	0.011	0.011	0.011	0.011	0.002	0.010	0.010
$\theta_0 = 0.2$	$\Pr(\theta = 0)$	0.788	0.796	0.789	0.760	0.261	0.847	0.955
	$\hat{\theta}$	0.046	0.042	0.042	0.048	0.155	0.010	0.000
	MSE	0.037	0.037	0.038	0.038	0.013	0.037	0.040
$\theta_0 = 0.3$	$\Pr(\theta = 0)$	0.522	0.543	0.519	0.459	0.084	0.671	0.934
	$\hat{\theta}$	0.147	0.144	0.149	0.168	0.273	0.141	0.000
	MSE	0.062	0.061	0.060	0.056	0.004	0.096	0.090
$\theta_0 = 0.4$	$\Pr(\theta = 0)$	0.210	0.236	0.203	0.157	0.005	0.450	0.902
	$\hat{\theta}$	0.344	0.335	0.345	0.364	0.404	0.285	0.000
	MSE	0.039	0.041	0.038	0.034	0.000	0.125	0.160
$\theta_0 = 0.5$	$\Pr(\theta = 0)$	0.056	0.067	0.050	0.031	0.014	0.263	0.870
	$\hat{\theta}$	0.494	0.489	0.497	0.506	0.484	0.359	0.000
	MSE	0.022	0.025	0.020	0.016	0.003	0.125	0.250

Normal example



Normal example



Normal example

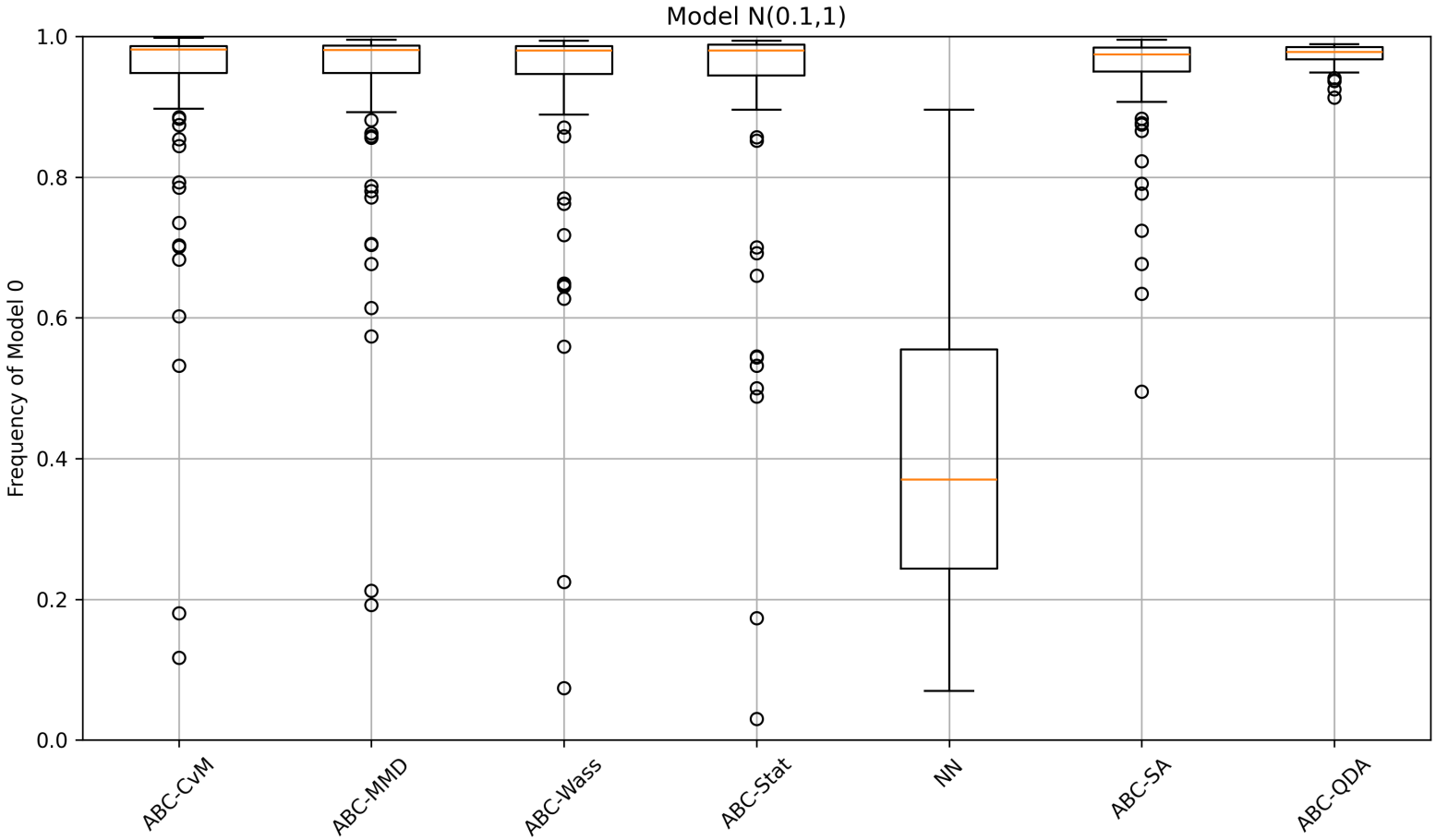


Figure: Boxplots of the estimated posterior probability of H_0 across 100 datasets generated from $N(0.1, 1)$.

Normal example

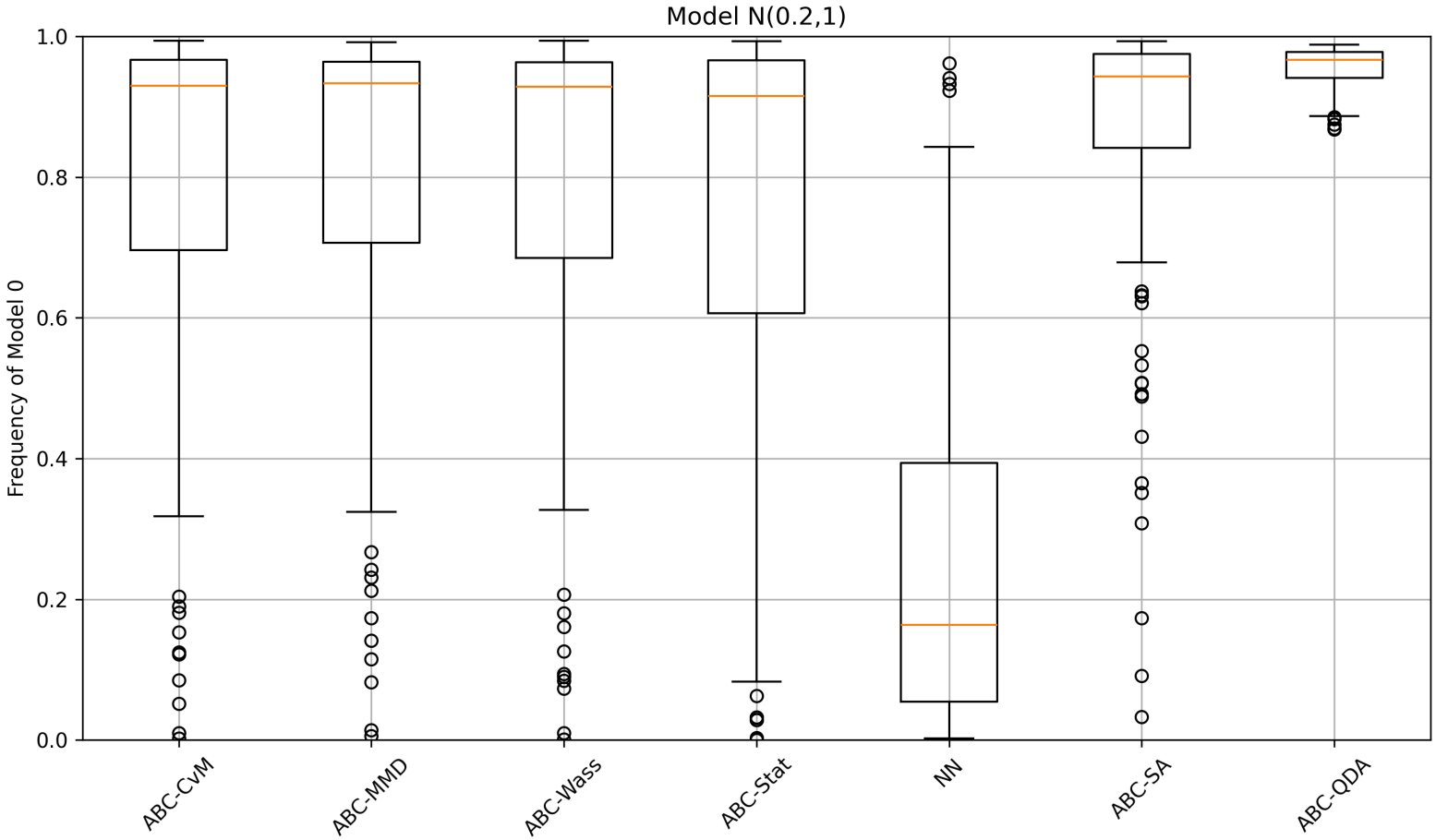


Figure: Boxplots of the estimated posterior probability of H_0 across 100 datasets generated from $N(0.2, 1)$.

Normal example

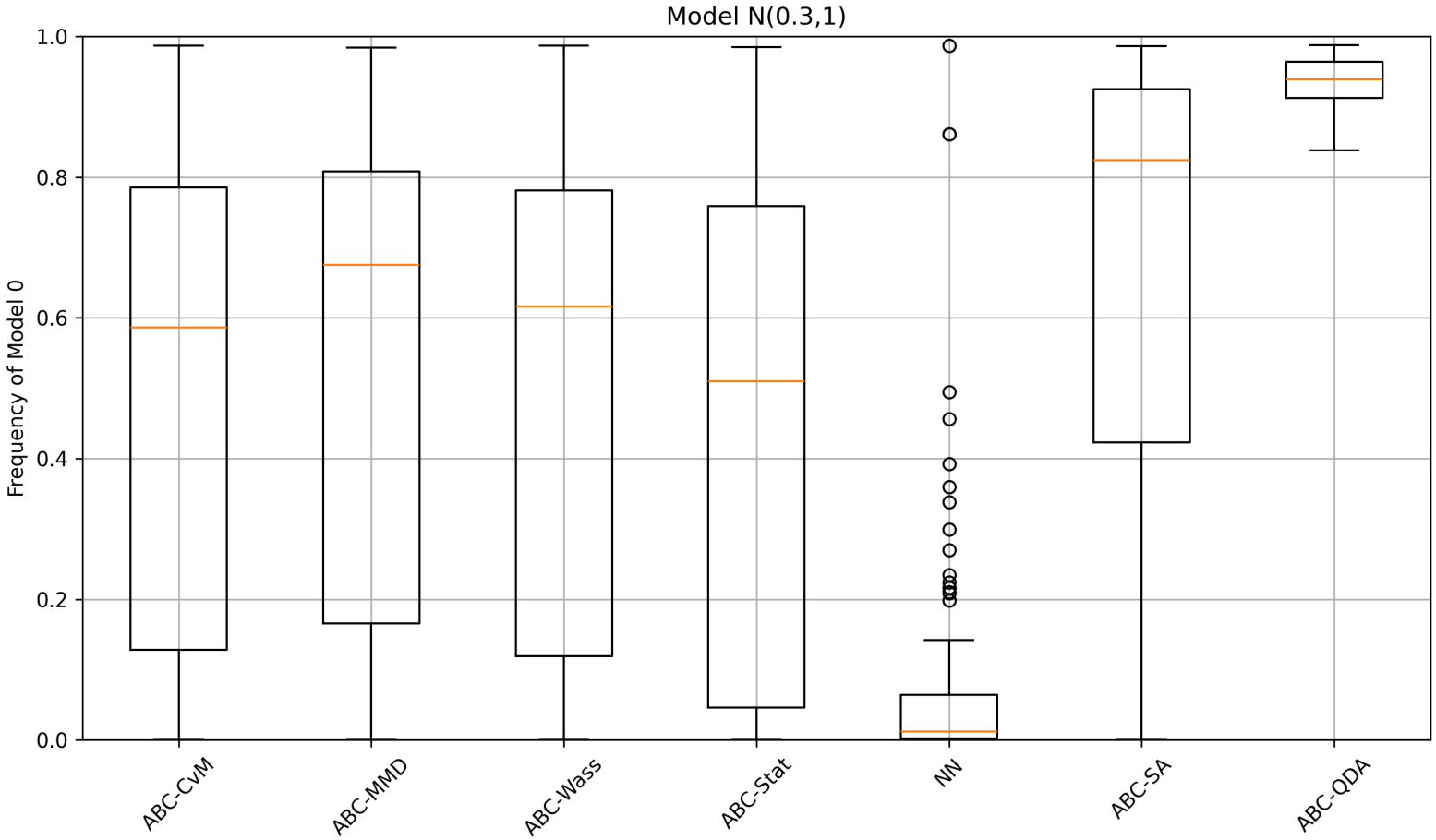


Figure: Boxplots of the estimated posterior probability of H_0 across 100 datasets generated from $N(0.3, 1)$.

Normal example

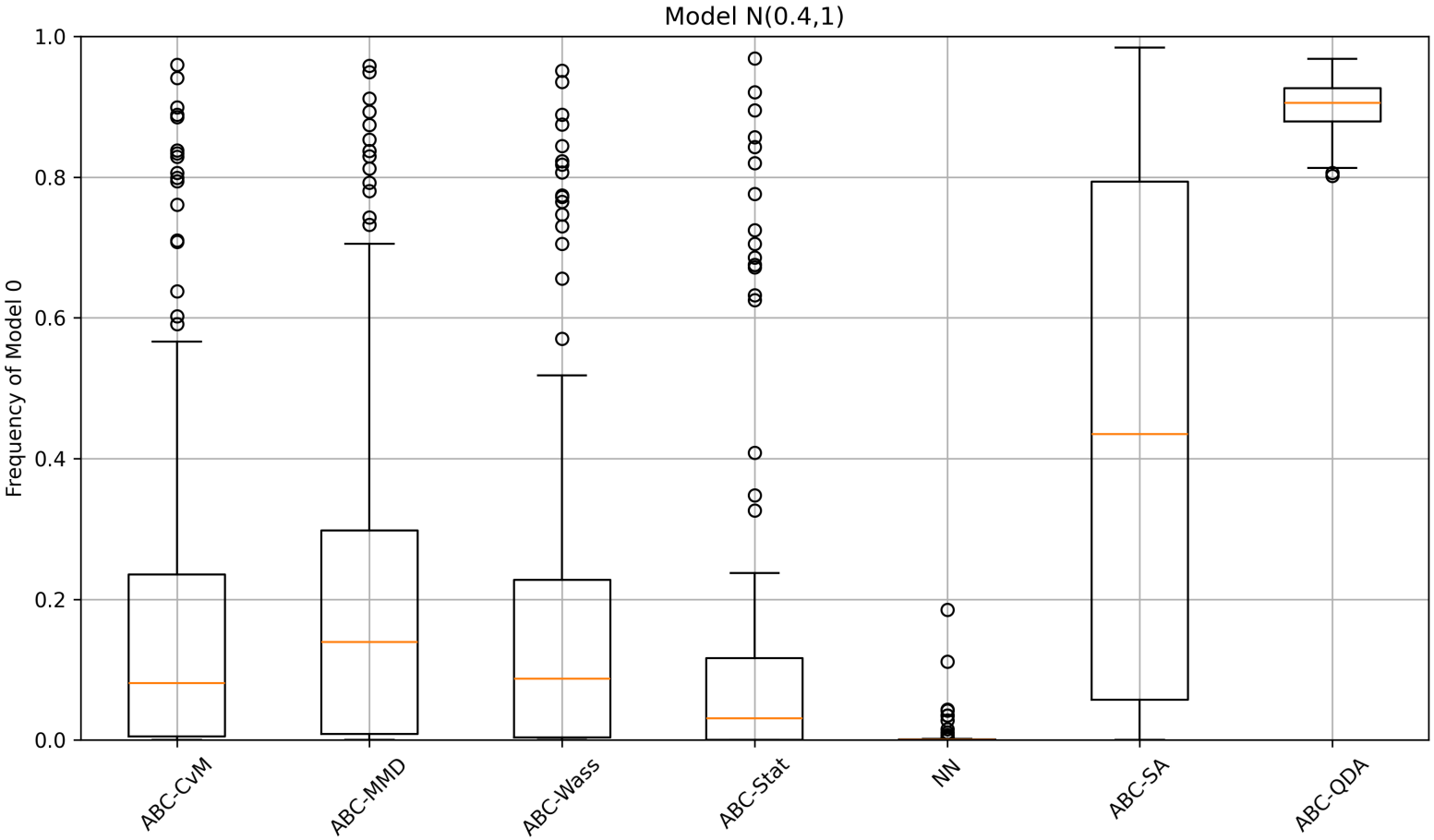


Figure: Boxplots of the estimated posterior probability of H_0 across 100 datasets generated from $N(0.4, 1)$.

Normal example

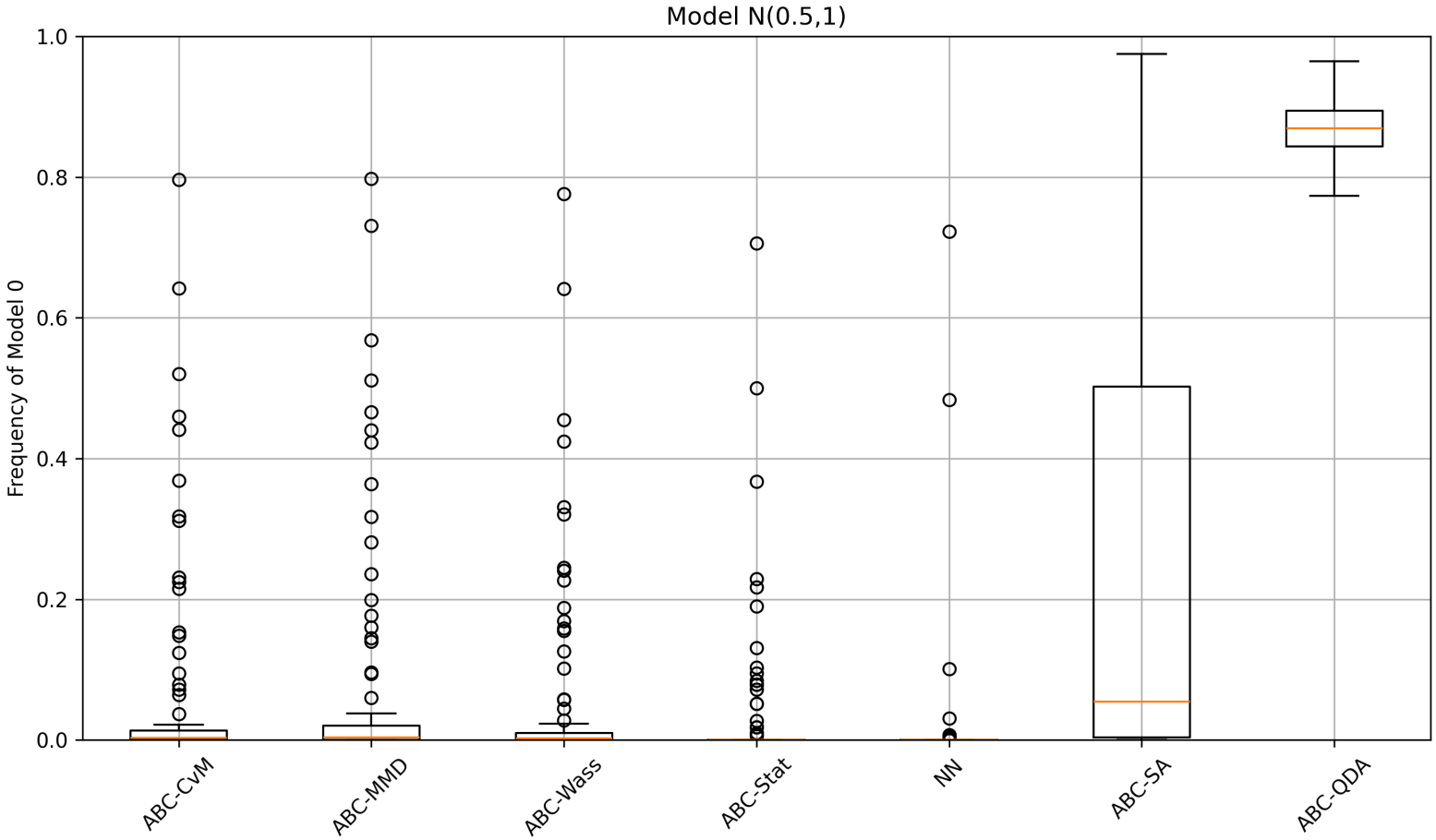


Figure: Boxplots of the estimated posterior probability of H_0 across 100 datasets generated from $N(0.5, 1)$.

Exponential family example

Consider three models from the exponential family.

- Model M_1: $y_i \overset{\text{i.i.d.}}{\sim} \mathscr{E}xp(\theta)$, $i = 1, \ldots, n$, with prior $\theta \sim \mathscr{E}xp(1)$.
- Model M_2: $y_i \overset{\text{i.i.d.}}{\sim} \mathscr{L}\mathscr{N}(\theta, 1)$, with prior $\theta \sim N(0,1)$.
- Model M_3: $y_i \overset{\text{i.i.d.}}{\sim} Ga(2, \theta)$, with prior $\theta \sim \mathscr{E}xp(1)$.

Marin et al. [2016] shows that no information is lost by replacing the full dataset with these summaries:

$$\boldsymbol{\eta}(\boldsymbol{y}) = \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} \log y_i, \sum_{i=1}^{n} \log^2 y_i \right).$$

Exponential family example

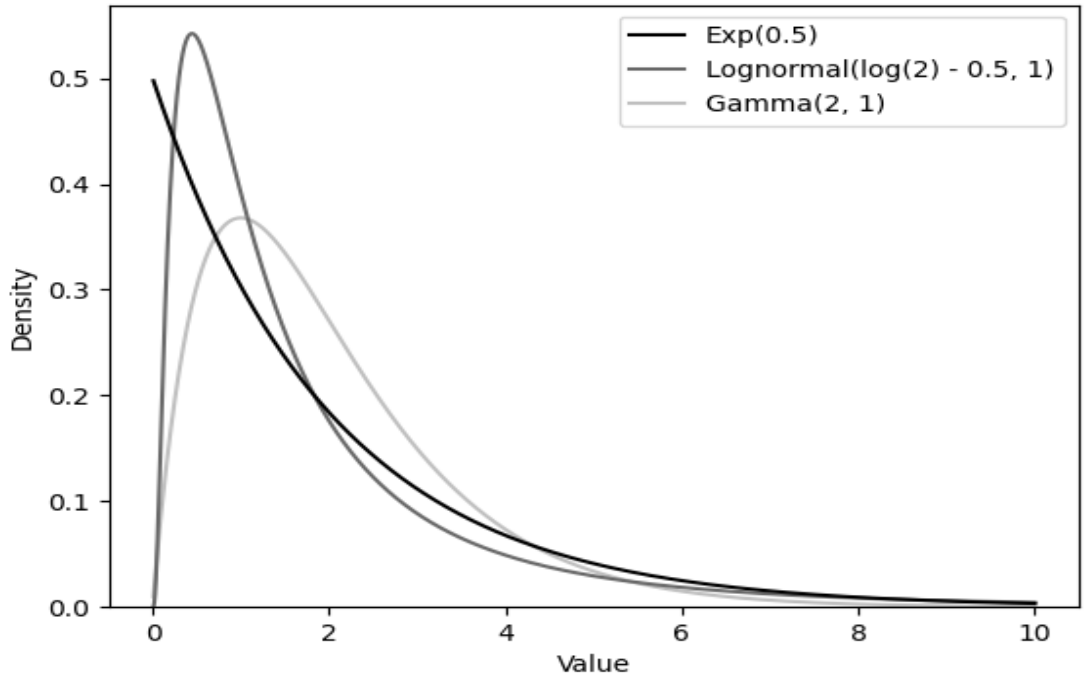


Figure: Comparison of the true data distributions for each model, with parameters set to $\theta = 1/2$ for M_1, $\theta = \log 2 - 1/2$ for M_2, and $\theta = 1$ for M_3. These values give each model a mean of 2.

Exponential family example

		ABC							ABC	
		CvM	**MMD**	**MMD (log)**	**Wass**	**Wass (log)**	**Stat**	**NN**	**SA**	**QDA**
$\mathscr{E}\text{xp}(0.5)$	$\Pr(M = M_1)$	0.883	0.836	0.953	0.850	0.948	0.952	0.554	0.772	0.922
	$\hat{\theta}$	0.512	0.506	0.511	0.515	0.514	0.513	0.454	0.510	0.522
	MSE	0.003	0.003	0.003	0.003	0.003	0.003	0.021	0.003	0.0040
$\mathscr{LN}(0.193,1)$	$\Pr(M = M_2)$	0.896	0.829	0.952	0.882	0.956	0.954	0.653	0.787	0.875
	$\hat{\theta}$	0.188	0.196	0.187	0.189	0.187	0.191	0.337	0.204	0.187
	MSE	0.009	0.008	0.008	0.008	0.009	0.010	0.034	0.013	0.140
$\mathscr{G}\text{a}(2,1)$	$\Pr(M = M_3)$	0.952	0.966	0.971	0.984	0.987	0.987	0.967	0.955	0.819
	$\hat{\theta}$	1.000	0.995	0.998	1.004	1.003	1.002	0.957	1.017	1.000
	MSE	0.004	0.004	0.004	0.004	0.004	0.004	0.008	0.008	0.275

Table: Results for the exponential family models across repetitions. For each model, we report the average posterior probability of selecting the correct model, the average estimate of the parameter θ, and the MSE across repetitions. Logarithm function is taken following Drovandi and Frazier [2022].

Exponential family example

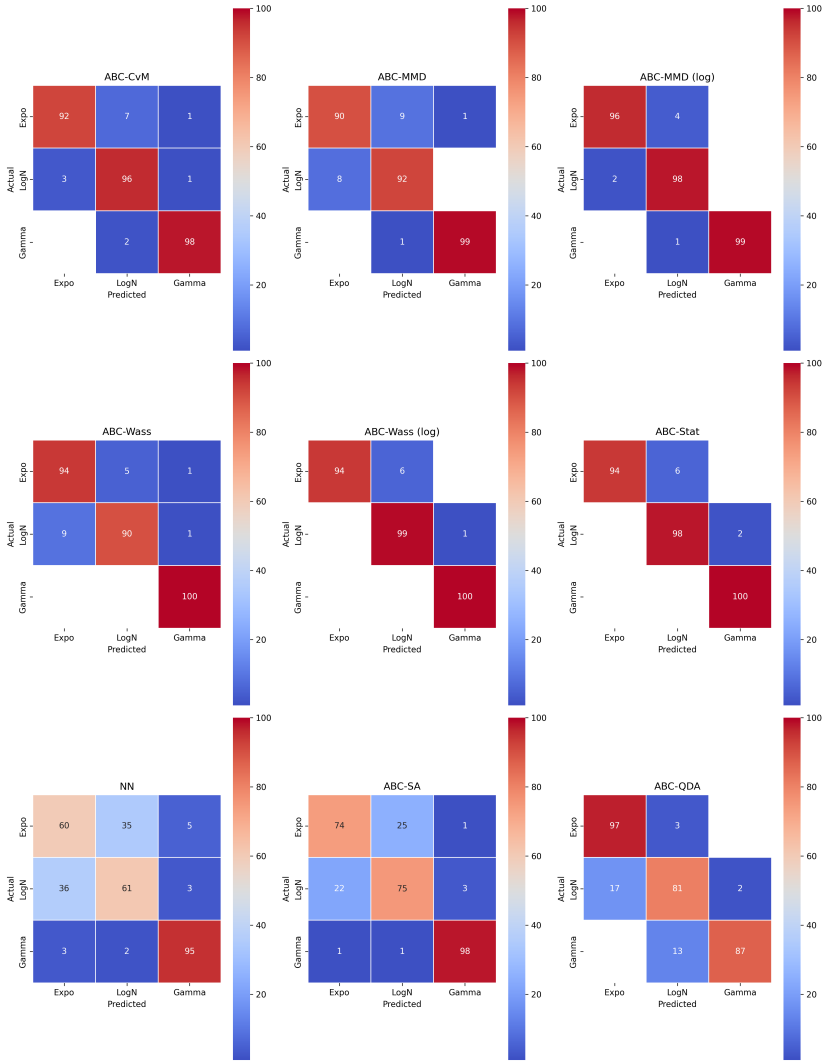


Figure: Confusion matrices for model selection in the exponential family example with $n = 100$.

Exponential family example

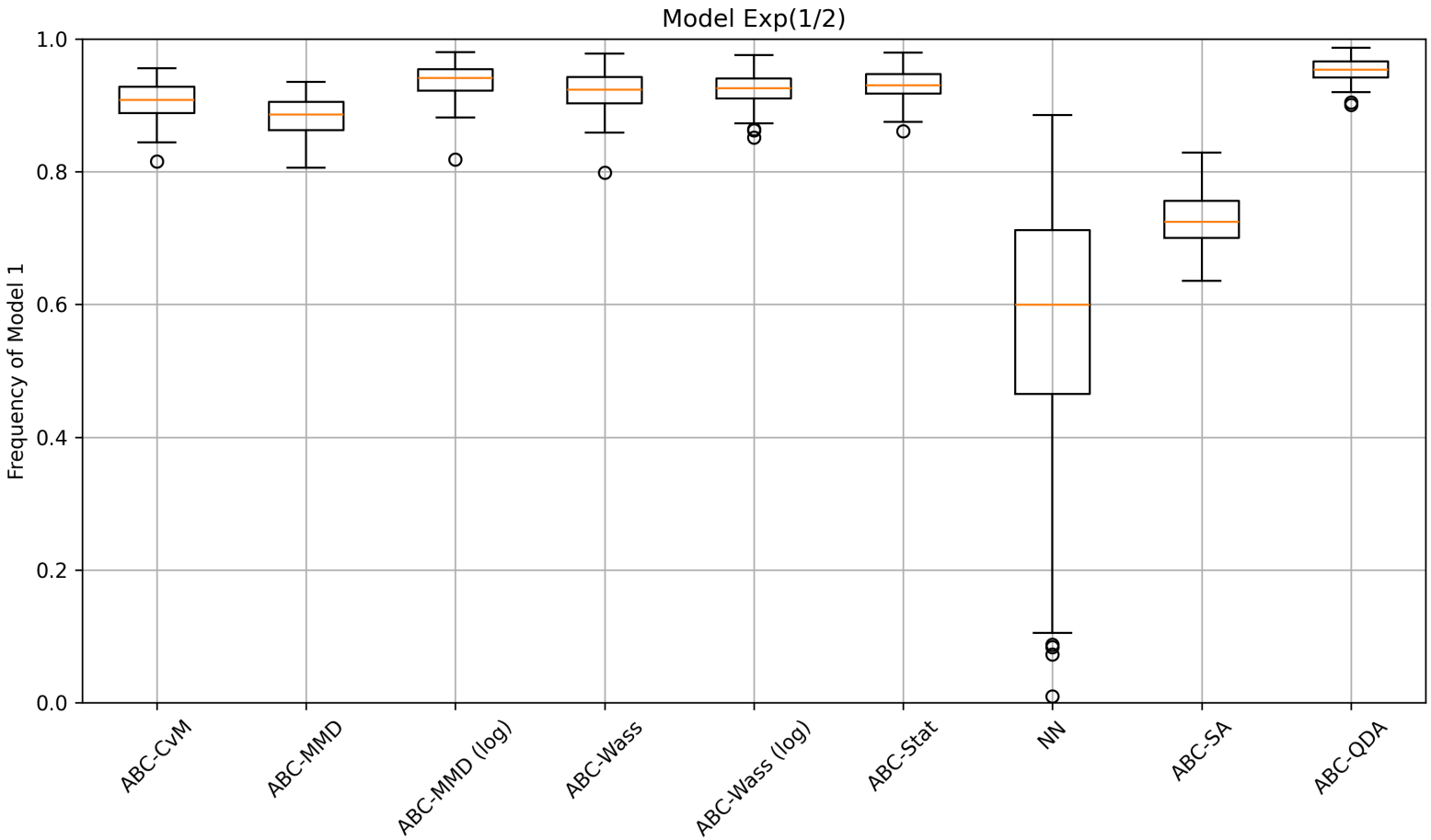


Figure: Boxplots of the estimated posterior probability of model M_1 across 100 datasets generated from $\mathscr{E}\text{xp}(1/2)$.

Exponential family example

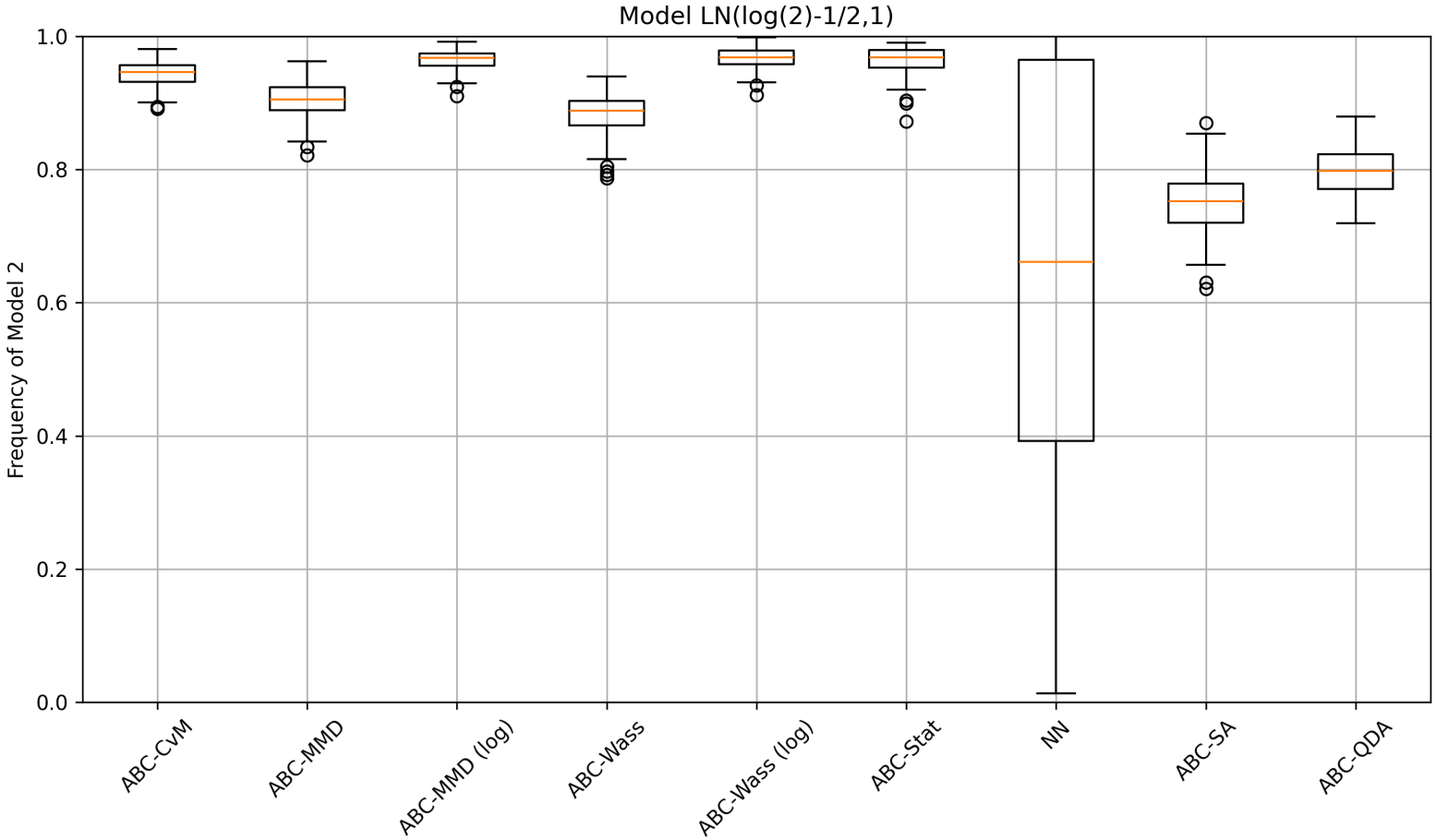


Figure: Boxplots of the estimated posterior probability of model M_2 across 100 datasets generated from $\mathscr{LN}(\log(2)-1/2,1)$.

Exponential family example

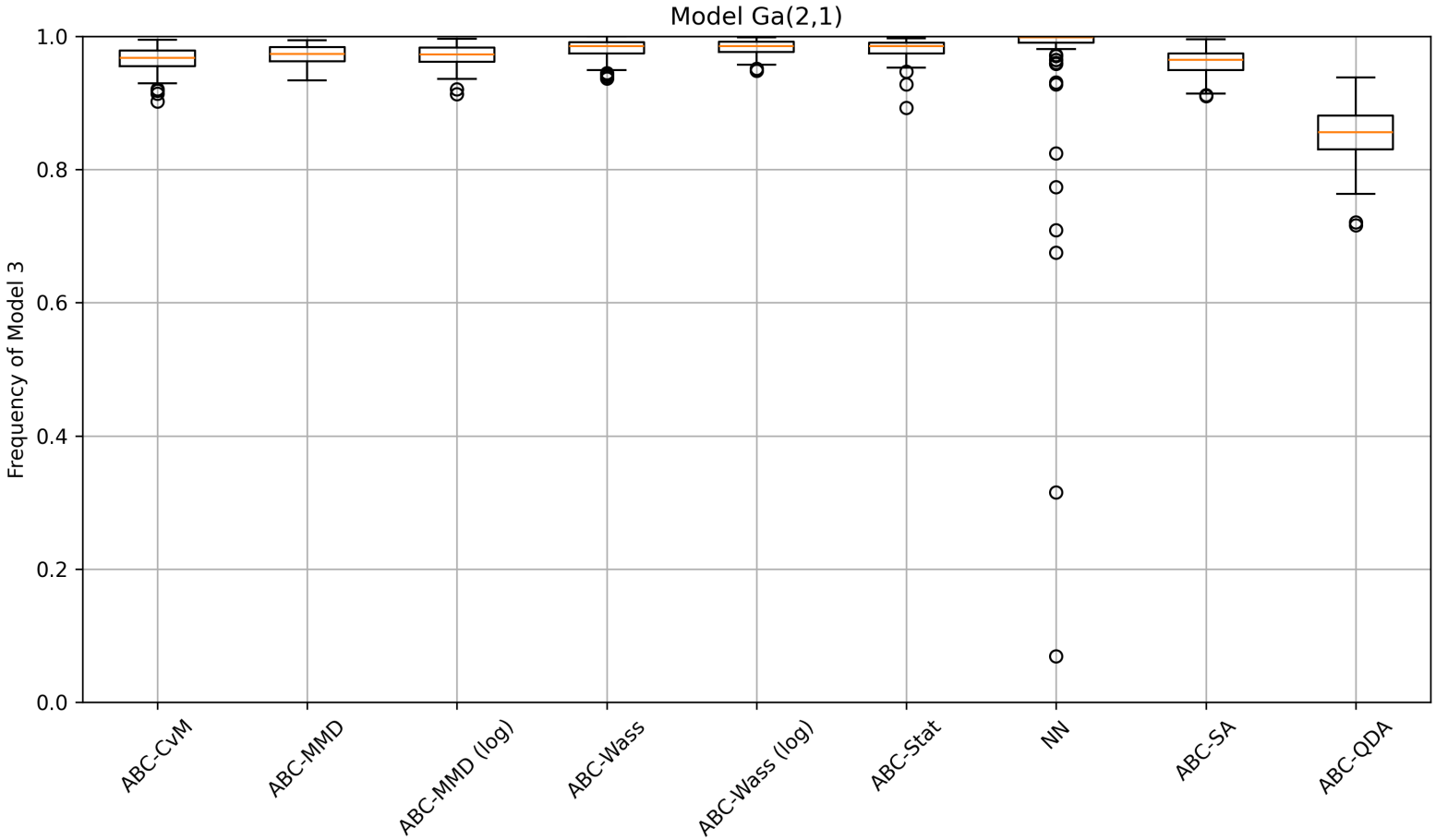
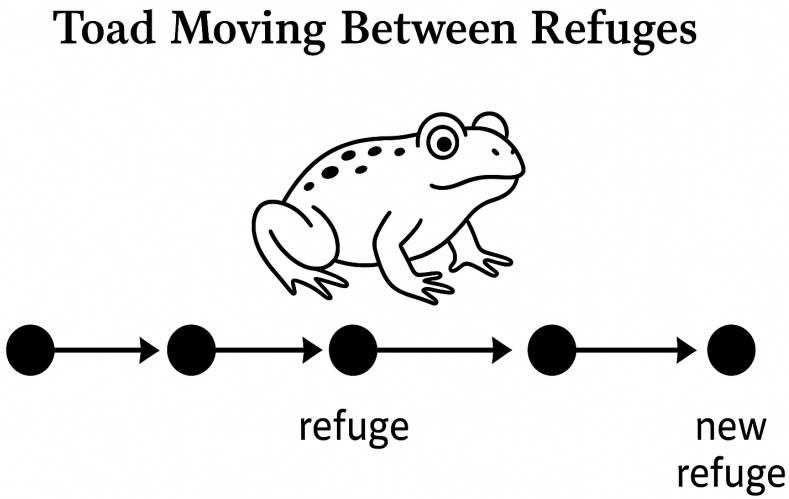


Figure: Boxplots of the estimated posterior probability of model M_3 across 100 datasets generated from $\mathscr{G}\mathrm{a}(2,1)$.

Real data example: Toad movement [Drovandi and Frazier, 2022]

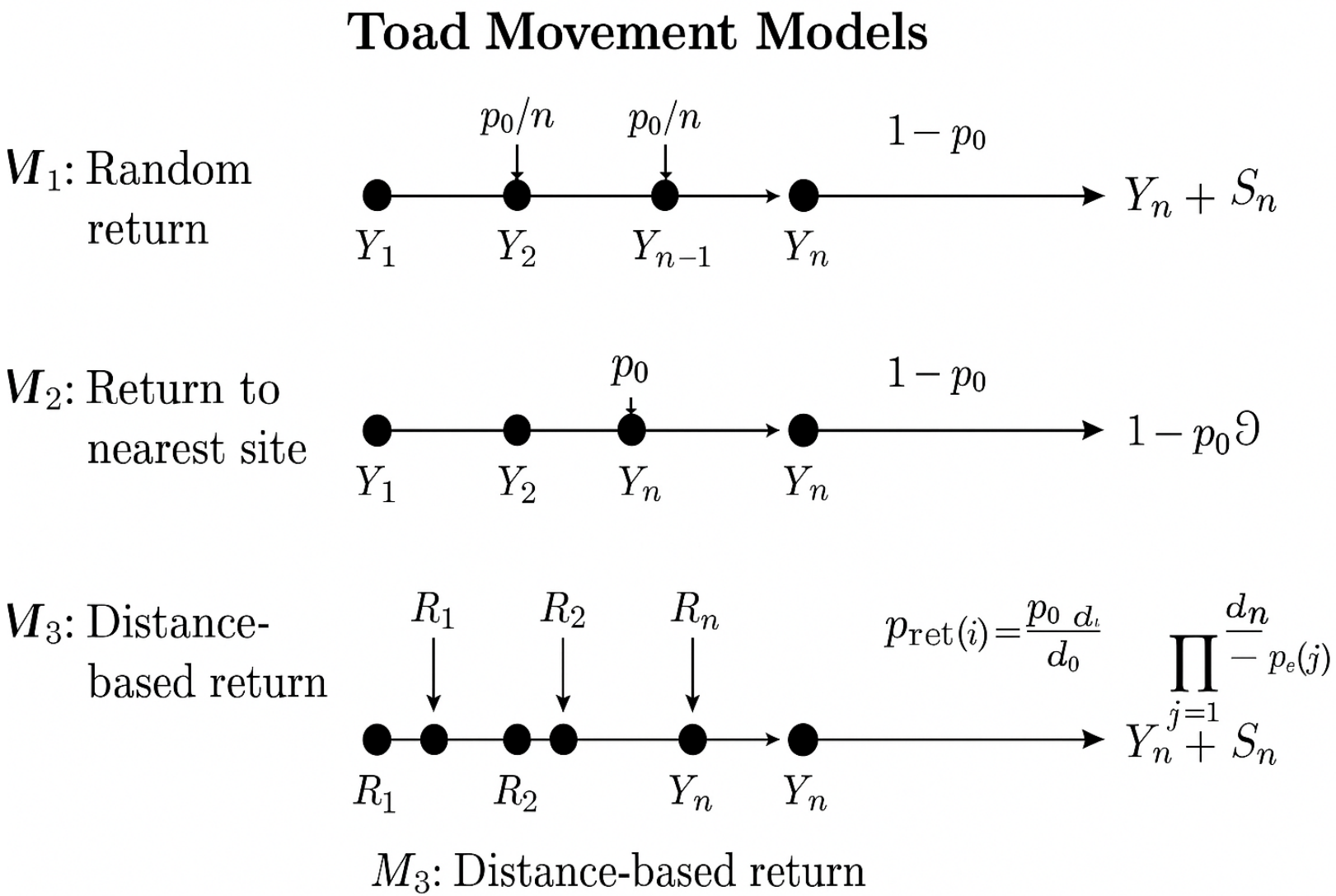


Following Marchand et al. [2017], the toad perform a displacement S_n from their last refuge location Y_n.

The toad either moves to a new location, $Y_{n+1} = Y_n + S_n$, or returns to a previous daytime refuge site (Y_i for some $i = 1, \ldots, n-1$).

The displacement follows a symmetric, zero-centered stable distribution, $S_n \sim S(\alpha, \gamma)$

Real data example: Toad movement



M_3: Distance-based return

Real data example: Toad movement

- Model M_1 is the random return model

$$Y_{n+1} = \begin{cases} Y_n + S_n & \text{with prob. } 1 - p_0, \\ Y_i & \text{with prob. } p_0/n \ \forall\, i = 1, \ldots, n. \end{cases}$$

- Model M_2 assumes that, when the toad returns, it always chooses the nearest refuge site.

$$Y_{n+1} = \begin{cases} Y_n + S_n & \text{with prob. } 1 - p_0, \\ \min\limits_{Y \in \{Y_1, \ldots, Y_n\}} |Y_{n+1} - Y| & \text{with prob. } p_0. \end{cases}$$

- Model M_3 is a distance-based return model

$$Y_{n+1} = \begin{cases} Y_n + S_n & \text{with prob. } \prod_{j=1}^{A_{n+1}} (1 - p_{ret(j)}), \\ R_i & \text{with prob. } p_i \ \forall\, i = 1, \ldots, A_{n+1}, \end{cases}$$

Real data example: Toad movement

	$\Pr(M = M_1)$	$\Pr(M = M_2)$	$\Pr(M = M_3)$
ABC-CvM	0.08	0.00	0.92
ABC-MMD	0.29	0.00	0.71
ABC-MMD (log)	0.07	0.00	0.93
ABC-Wass	0.14	0.00	0.86
ABC-Wass (log)	0.00	0.00	1.00
ABC-Stat	0.40	0.00	0.60
ABC-Stat [Marchand et al., 2017]	0.15	0.00	0.85
NN	0.32	0.02	0.66
ABC-SA	0.36	0.00	0.64
ABC-QDA	0.13	0.09	0.79

Table: Estimated posterior probabilities of the three toad movement models, comparing different methods with the original results of Marchand et al. [2017].

Conclusions

This work has investigated ABC as a tool for model selection in settings with intractable likelihoods.

- full data ABC approaches yield posterior distributions closely matching ABC with sufficient statistics (when they are sufficient for model choice problems) or provides high support to the correct model
- the Wasserstein distance consistently produced the most accurate model selection results
- theoretical results establish both the consistency of ABC-Wass in selecting the correct model and its robustness to model misspecification
- future work will extend these results by leveraging the general framework of Legramanti et al. [2025]
- curse of dimensionality?

References I

Chris Barnes, Sarah Filippi, Michael Stumpf, and Thomas Thorne. Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, 22:1181—-1197, 11 2012.

Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. Approximate Bayesian Computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, February 2019. ISSN 1467-9868.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. New insights into approximate bayesian computation. *Annales de l'Institut Henri Poincaré*, 51(1):376–403, 2015.

Michael GB Blum and Olivier François. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*, 20:63–73, 2010.

C. Drovandi and D.F. Frazier. A comparison of likelihood-free methods with and without summary statistics. *Statistics and Computing*, 32(42), 2022.

Florence Forbes, Hien Duy Nguyen, TrungTin Nguyen, and Julyan Arbel. Summary statistics and discrepancy measures for approximate bayesian computation via surrogate posteriors. *Statistics and Computing*, 32(5):85, 2022.

David T. Frazier. Robust and efficient approximate Bayesian computation: A minimum distance approach. arXiv:2006.14126, 2020.

References II

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013.

Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28:411–425, 2018.

Bai Jiang, Tung-Yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618, 2017.

Sirio Legramanti, Daniele Durante, and Pierre Alquier. Concentration of discrepancy-based ABC via Rademacher complexity. 2025.

Philippe Marchand, Morgan Boenke, and David M. Green. A stochastic movement model reproduces patterns of site fidelity and long-distance dispersal in a population of Fowler's toads (Anaxyrus fowleri). *Ecological Modelling*, 360:63–69, 2017. ISSN 0304-3800.

Jean-Michel Marin, Natesh S. Pillai, Christian P. Robert, and Judith Rousseau. Relevant Statistics for Bayesian Model Choice. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(5):833–859, 12 2013.

References III

Jean-Michel Marin, Pierre Pudlo, Arnaud Estoup, and Christian P. Robert. *Likelihood-free Model Choice*, chapter 6. CRC Press Taylor & Francis Group, 2016.

Gael M. Martin, David T. Frazier, and Christian P. Robert. Approximating Bayes in the 21st century. *Statistical Science*, 39(1):20–45, 2024.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*, pages 39–68. Springer International Publishing, 2022.

Hien D. Nguyen, Julyan Arbel, Hongliang Lü, and Florence Forbes. Approximate bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698, 2020.

Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR*, 51:398–407, 2016.

Dennis Prangle, Paul Fearnhead, Murray P. Cox, Patrick J. Biggs, and Nigel P. French. Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, 13(1):67–82, 2014.

References IV

C.P. Robert, J. Cornuet, J. Marin, and N.S. Pillai. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117, 2011.

Sara Sheehan and Yun S Song. Deep learning for population genetic inference. *PLoS Computational Biology*, 12(3):e1004845, 2016.