13. Bayesian Inference for Extreme Value Modelling Supplement

Alec Stephenson

1.8 Model Diagnostics

Any analysis should include some check of the adequacy of the fit of the model to the data, and of the plausibility of the model for the purposes for which it will be used. In a Bayesian context, the model refers to both the prior distribution $f(\theta)$ and the likelihood model $f(y|\theta)$. In this section we briefly discuss aspects of model checking, sensitivity analysis, and model comparisons. The description of posterior predictive checking is based on Gelman *et al.* (2013). A balanced discussion of the advantages and disadvantages of this approach is given by Bayarri and Berger (1999, 2000). Further examples are given in Gelman *et al.* (1996).

In practice, additional information is often available that is not included formally in the likelihood or the prior distribution. If this information suggests that posterior inferences are false, then more effort should be made to incorporate this information within the model. We can perform informal diagnostic procedures by comparing posterior distributions and posterior predictive distributions with aspects of reality that are not captured by the model. If there are any discrepancies, the model should be extended to include these aspects.

A more formal diagnostic procedure compares the posterior predictive distribution to the data that have been observed. This is known as posterior predictive checking. This is a simple technique for checking whether the observed data y looks plausible under the posterior predictive distribution $f(\tilde{y}|y)$. If the model is appropriate, then data generated using $f(\tilde{y}|y)$ should be similar to the observed data y. We therefore simulate samples from the posterior predictive distribution $f(\tilde{y}|y)$. These samples are then compared to the original data. Systematic discrepancies between the samples and the data correspond to features that are poorly fitted by the model.

Suppose we simulate $\tilde{y}^1, \tilde{y}^2, \tilde{y}^3, \ldots, \tilde{y}^N$ from the posterior predictive density $f(\tilde{y}|y)$. This can be achieved using equation (13.3) in Section 13.1.1. Markov chain Monte Carlo output produces draws $\theta^1, \theta^2, \theta^3, \ldots, \theta^N$ from the posterior distribution $f(\theta|y)$, and we can then draw the vector $\tilde{y}^t$ from $f(\tilde{y}|\theta^t)$ for each $t = 1, \ldots, N$. The length of each vector $\tilde{y}^t$ is the same as the length of y.

It can be difficult to compare the N posterior predictive samples $\tilde{y}^t$ to the actual data y using only graphical methods. Instead, we can define some function of the data $T(\cdot)$. We can then calculate the number of samples from the posterior predictive distribution for which the test statistic $T(\cdot)$ is greater than the value calculated for the actual data. We define p to be the proportion of the simulations for which $T(\tilde{y}^t) > T(y)$. If the value of p

is close to zero or one, the test statistic $T(\cdot)$ corresponds to a feature that is poorly fitted by the model. The test statistic $T(\cdot)$ should be chosen to reflect aspects of the model that are relevant to the purposes to which the inference will be applied. In particular, $T(y) = \max(y)$ may be of importance for extreme value models. We can also let the test statistic depend on the parameters θ, and p can then be defined as the proportion of the simulations for which $T(\tilde{y}^t, \theta^t) > T(y, \theta^t)$.

It is often the case that more than one model provides an adequate fit to the data. Sensitivity analysis determines by what extent posterior inferences change when alternative models are used. Alternative models may differ in the likelihood $f(y|\theta)$, or in terms of the prior specification $f(\theta)$. The basic method of sensitivity analysis is to fit several models to the same problem. Posterior inferences from each model can then be compared. Posterior inferences for the GEV model will typically include marginal posterior distributions of the parameters (μ, σ, ξ) and posterior distributions of quantiles. The sensitivity of the marginal posterior density of the shape parameter ξ is often of particular interest.

If we have a large amount of data we could test predictive accuracy by fitting our model to only e.g. two-thirds of the data y_{train}, randomly selected, and use the remaining third y_{test} as a test set. We could then derive a score of predictive accuracy by using y_{test} to evaluate a suitable metric. For example, if we have N draws $\theta^1, \theta^2, \theta^3, \ldots, \theta^N$ from the posterior distribution $f(\theta|y_{train})$ we could score each value y^* in y_{test} using

$$\log\left(\frac{1}{N}\sum_{t=1}^{N} f(y^*|\theta^t)\right).$$

The scores can be summed over y_{test} to give a measure of predictive accuracy. This measure can then be used across different models for the purposes of model comparison, with larger values indicating better models. Unfortunately we often do not have large amounts of data and so it is necessary to use all data for model fitting. However we can extend the above approach in an obvious manner using cross-validation techniques. The cross-validation approach is often useful in simpler models but can be computationally demanding.

An alternative approach is to employ information criteria. Various information criteria have been proposed for comparing different models. These criteria can be regarded as a simplified approach for the evaluation of predictive accuracy, and they are far less computationally demanding that the cross-validation methodology outlined above. With N draws $\theta^1, \theta^2, \theta^3, \ldots, \theta^N$ from $f(\theta|y)$, one basic approach is to use

$$\log f(y|\bar{\theta}) - p^* \tag{1}$$

where p^* represents the effective number of parameters in the model, and $\bar{\theta}$ is the estimated posterior mean. Different methods have been proposed for estimating p^*. One popular method is the Deviance Information Criterion (DIC) of Spiegelhalter *et al.* (2002), which uses

$$p^* = 2\left(\log f(y|\bar{\theta}) - \frac{1}{N}\sum_{t=1}^{N}\log f(y|\theta^t)\right).$$

The DIC is defined on the deviance scale, so that (1) is multiplied by minus two, with smaller DIC values indicating better models. We therefore have

$$\text{DIC} = 2\log f(y|\bar{\theta}) - \frac{4}{N}\sum_{t=1}^{N}\log f(y|\theta^t). \tag{2}$$

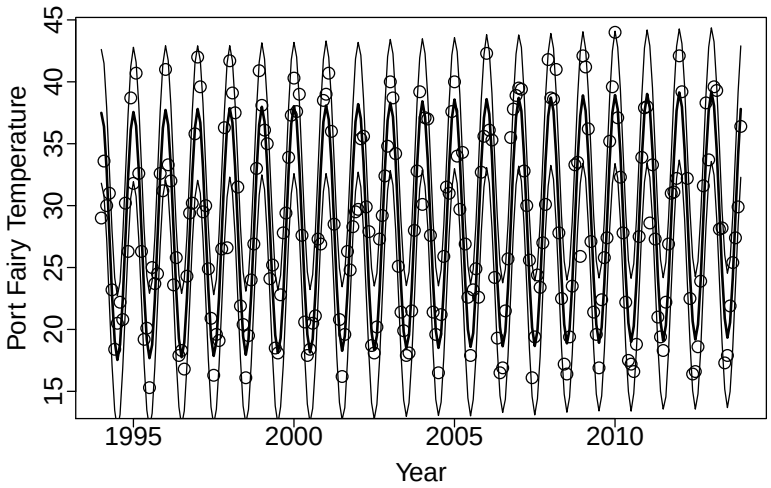

Figure 1: Maximum monthly temperatures at the Port Fairy weather station for the period 1995-2013. The curves give the median, the 5% quantile, and the 95% quantile from the predictive posterior distribution of the model fitted in this section.

The Markov chain Monte Carlo algorithms in Sections 13.1.3 and 13.1.4 can be adjusted so that $\log f(y|\theta^t)$ is computed and stored at the end of each iteration in the main for loop. The DIC is then simple to calculate.

An alternative criterion that has been proposed more recently is the Widely Applicable Information Criterion (WAIC) of Watanabe (2010), given by

$$\text{WAIC} = \sum_{i=1}^{n} \left\{ 2 \log \left(\frac{1}{N} \sum_{t=1}^{N} f(y_i|\theta^t) \right) - \frac{4}{N} \sum_{t=1}^{N} \log f(y_i|\theta^t) \right\}, \tag{3}$$

for data $y = (y_1, \ldots, y_n)$. The simplest approach to the calculation of the WAIC is to adjust the Markov chain Monte Carlo algorithm to compute $\log f(y_i|\theta^t)$ for every $i = 1, \ldots, n$ at the end of each iteration in the main for loop. For both the DIC and the WAIC there exist alternative proposals for estimating the effective number of parameters. See Gelman *et al.* (2014) for a comparative review.

2.2 Bivariate Extreme Value Modelling

The data we model here consist of monthly temperature maxima recorded at two different weather stations in Australia over the period from January 1994 to December 2013. The first weather station ($j = 1$) is located at Port Fairy, which is on the coastline, approximately 300 kilometers to the West of Melbourne. The second weather station ($j = 2$) is located at Melbourne Airport, which is approximately 25 kilometers North-West of the centre of Melbourne. Dependence between the temperature maxima at the two sites is to be expected. The Port Fairy data is shown in Figure 1. We clearly need to account for the seasonality in the monthly maxima, otherwise the level of dependence between the sites will be over-estimated.

Bivariate extreme value distributions are a natural way to model this type of data. The univariate marginal distributions are generalized extreme value, with parameters (μ_j, σ_j, ξ_j) on the jth margin. There are various ways to parameterize the dependence

structure of the distribution. We use the logistic dependence structure, which contains a single parameter $0 < \gamma \leq 1$. When $\gamma = 1$ the two margins are independent. The dependence increases as γ decreases. The distribution function is given by

$$F(x_1, x_2) = \exp\left\{-\left(y_1^{1/\gamma} + y_2^{1/\gamma}\right)^{\gamma}\right\}, \tag{4}$$

where

$$y_j = y_j(x_j) = \left(1 + \xi_j \left(\frac{x_j - \mu_j}{\sigma_j}\right)\right)_+^{-1/\xi_j} \tag{5}$$

for $j = 1, 2$, and where $h_+ = \max\{h, 0\}$. For alternative dependence structures, see e.g. Tawn (1988) and Boldi and Davison (2007).

We let the location parameters of each site depend on the month t, accounting for a possible linear trend. We also account for seasonality by fitting a single sinusoid with a frequency of one year, where we estimate both the amplitude and the phase. We could also investigate similar temporal models for the marginal scale and shape parameters, but to keep the model simple we take these parameters to be fixed across time. This gives the model for the jth margin as

$$\begin{aligned}
\mu_{j,t} &= \beta_j^{(0)} + \beta_j^{(1)} t + \beta_j^{(2)} \cos(2\pi t/12 + \beta_j^{(3)}) \\
\sigma_{j,t} &= \sigma_j, \\
\xi_{j,t} &= \xi_j,
\end{aligned}$$

for $j = 1, 2$. The model therefore has 12 parameters in total to be estimated: the six parameters $(\beta_j^{(0)}, \beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)}, \sigma_j, \xi_j)$ on each of the two margins, and the dependence parameter $0 < \gamma \leq 1$. We take the time origin $t = 0$ to be the month of January 2010.

Our first step is to construct the prior. We specify a weakly informative prior on $(\beta_j^{(0)}, \sigma_j, \xi_j)$ using the method given in Section 13.1.6.2, which employs beta distributions for probability ratios. The parameter $\beta_j^{(0)}$ represents the generalized extreme value location in the autumn and spring seasons when there is no linear trend. Using general climate information we specify a beta(2,2) prior distribution for the probability that an autumn monthly maxima exceeds 28 degrees Celsius. Roughly half of the monthly maxima that exceed 28 degrees will also exceed 30 degrees, and roughly half of the monthly maxima that exceed 30 degrees will also exceed 32 degrees. Using the notation of Section 13.1.6.2, this gives a prior distribution with hyperparameters $\alpha = (2, 1, 0.5, 0.5)$, which is applied independently to both margins.

For convenience we reparameterize our weakly informative prior to $(\beta_j^{(0)}, \nu_j, \xi_j)$, where $\nu_j = \log(\sigma_j)$. This gives the following code. The term `jac` is the Jacobian $J(\theta)$ from Section 13.1.6.2. The inclusion of `nu` in the returned value `ld + jac + nu` is due to the logarithmic transformation of the scale parameter.

```
log_prior <- function(mu, nu, xi) {
  quant <- c(28,30,32)
  alpha <- c(2,1,0.5,0.5)
  z <- 1 + xi * (quant - mu) / exp(nu)
  if(any(z <= 0)) return(-Inf)
  z <- z^(-1/xi)
  pd <- -diff(c(1, 1-exp(-z), 0))
  if(any(pd <= 0)) return(-Inf)
```

```
  ld <- sum((alpha-1) * log(pd))

  jac <- (z[1]*z[2])^(-xi) * log(z[2]/z[1]) -
    (z[1]*z[3])^(-xi) * log(z[3]/z[1]) +
    (z[2]*z[3])^(-xi) * log(z[3]/z[2])
  jac <- log(abs(jac)) - 2*nu - log(xi^2)
  jac <- jac + (1 + xi) * sum(log(z)) - sum(z)

  ld + jac + nu
}
```

For the amplitudes $\beta_j^{(2)}$ and the linear trends $\beta_j^{(1)}$ we specify vague prior normal distributions. We expect the phase parameters $\beta_j^{(3)}$ to be fairly close to zero since the monthly maxima are largest in the Australian summer. We therefore take the phase prior distributions as normal with zero mean and a standard deviation of 0.5. For the dependence parameter γ we take a prior uniform distribution on the interval $[0, 1]$.

Putting the above elements together yields the `log_post` function as given below. The `dbvevd` function can be used to calculate the density of the bivariate extreme value distribution with logistic dependence structure. The function arguments `mp1` and `mp2` correspond to the parameter vectors $(\beta_j^{(0)}, \beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)}, \sigma_j, \xi_j)$ for $j = 1, 2$ respectively. The argument `dep` corresponds to the dependence parameter γ.

```
log_post <- function(mp1, mp2, dep, data) {
  tt <- (1:nrow(data) - 193)/12
  if(dep <= 0 || dep > 1) return(-Inf)
  m1v <- mp1[1] + mp1[2] * tt + mp1[3] * cos(2*pi*tt + mp1[4])
  m1v <- cbind(m1v, exp(mp1[5]), mp1[6])
  m2v <- mp2[1] + mp2[2] * tt + mp2[3] * cos(2*pi*tt + mp2[4])
  m2v <- cbind(m2v, exp(mp2[5]), mp2[6])

  llhd <- dbvevd(data, dep, mar1 = m1v, mar2 = m2v, log = TRUE)
  llhd <- sum(llhd, na.rm = TRUE)
  lprior1 <- sum(dnorm(mp1[2:4], sd = c(10,10,0.5), log = TRUE))
  lprior1 <- lprior1 + log_prior(mp1[1], mp1[5], mp1[6])
  lprior2 <- sum(dnorm(mp2[2:4], sd = c(10,10,0.5), log = TRUE))
  lprior2 <- lprior2 + log_prior(mp2[1], mp2[5], mp2[6])
  lprior1 + lprior2 + llhd
}
```

There are a small number of missing values in the data. For convenience, if it is missing at either site we simply ignore the corresponding month. We also rescale the time vector so that it represents years rather than months. Define $\beta_j = (\beta_j^{(0)}, \beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)})$ for $j = 1, 2$. Numerical optimization suggests that the posterior density has a global optimum at $\hat{\beta}_1 = (27.84, 0.09, 10.03, -0.03)$, $\hat{\beta}_2 = (27.76, 0.09, 10.52, 0.04)$, $\hat{\nu}_1 = 1.19$, $\hat{\nu}_2 = 0.92$, $\hat{\xi}_1 = -0.34$, $\hat{\xi}_2 = -0.13$ and $\hat{\gamma} = 0.58$.

The Markov chain Monte Carlo simulation proceeds as in the previous data example of Section 13.2.1. The autocorrelations and cross-correlations of the Markov chain simulations are weaker here, and therefore the chain has better mixing properties than in Section 13.2.1. We generate four different chains of length 4 000, combining the last half of each chain to use the combined 8 000 iterations for estimation purposes.

	0.025	0.25	Median	0.75	0.975
$\beta_1^{(0)}$	27.30	27.67	27.84	28.03	28.36
$\beta_1^{(1)}$	0.03	0.07	0.09	0.11	0.15
$\beta_1^{(2)}$	9.41	9.80	10.00	10.19	10.58
$\beta_1^{(3)}$	-0.08	-0.05	-0.03	-0.01	0.02
ν_1	1.11	1.17	1.20	1.23	1.29
ξ_1	-0.39	-0.36	-0.34	-0.32	-0.28
$\beta_2^{(0)}$	27.33	27.61	27.76	27.91	28.26
$\beta_2^{(1)}$	0.04	0.08	0.10	0.11	0.15
$\beta_2^{(2)}$	10.00	10.31	10.49	10.67	11.05
$\beta_2^{(3)}$	-0.01	0.02	0.04	0.05	0.08
ν_2	0.83	0.89	0.93	0.96	1.04
ξ_2	-0.22	-0.16	-0.12	-0.09	-0.02
γ	0.52	0.56	0.58	0.61	0.66

Table 1: Estimated quantiles for marginal posterior distributions for each of the thirteen model parameters.

We make a slight amendment to the usual simulation algorithm to account for the dependence parameter constraint $0 < \gamma \leq 1$. We use a logit-normal proposal distribution for γ. The logit-normal proposal distribution is asymmetric, and therefore the proposal density term $p_{y,j}(\cdot|\cdot)$, as defined in Section 13.1.4, does not cancel in the acceptance ratio. The code below represents the updating step for the dependence parameter, which is the last parameter to be updated. The calculation of the acceptance ratio r includes both the posterior density ratio and the proposal density ratio.

```
propmn13 <- log(out[t,13]/(1 - out[t,13]))
prop13 <- rnorm(1, mean = propmn13, propsd[13])
prop13 <- exp(prop13)/(1 + exp(prop13))
lpost_prop <- log_post(out[t+1,1:6],out[t+1,7:12],prop13,data)
r <- exp(lpost_prop - lpost_old)
r <- r * (prop13*(1-prop13)) / (out[t,13]*(1-out[t,13]))
if(r > runif(1)) {
  out[t+1,13] <- prop13
  lpost_old <- lpost_prop
}
else out[t+1,13] <- out[t,13]
```

Estimated quantiles for the marginal posterior distributions for each of the model parameters are given in Table 1. It can be seen from the parameters $\beta_1^{(2)}$ and $\beta_2^{(2)}$ that there is a strong seasonal component at both Port Fairy and Melbourne Airport. The dependence between the sites is fairly strong, with the 95% Bayesian confidence interval for γ equal to $(0.52, 0.66)$. The is also some evidence of an increasing trend on both margins, with an increase in the location parameters of around 0.1 degrees Celsius per year over the 10 year period.

To investigate the increasing trends in more detail we fit three alternative models and then use the DIC and WAIC criteria as given in equations (2) and (3) respectively to perform model comparison. For the first model we specify $\beta_1^{(1)} = 0$, so that no linear trend is modelled at Port Fairy. For the second model we specify $\beta_2^{(1)} = 0$, so that no linear trend

is modelled at Melbourne Airport. For the third model we specify $\beta_1^{(1)} = \beta_2^{(1)} = 0$, so that no linear trend is modelled at either site.

Table 2 gives the information criteria for these models. For each model we generate four chains with starting values that are dispersed relative to the target distribution, and we evaluate the information criteria separately for each of the four chains. We discard the first 2 000 iterations of each chain and therefore the criteria are calculated using the remaining 2 000 iterations. The variability of the information criteria across the four chains is reassuringly small. Table 2 corroborates the evidence of an increasing trend on both margins, with the DIC and WAIC values being smaller for the model that includes the linear trend components.

	Run	Full	$\beta_1^{(1)} = 0$	$\beta_2^{(1)} = 0$	$\beta_1^{(1)} = \beta_2^{(1)} = 0$
DIC	1	50.48	56.21	58.89	58.35
	2	50.68	56.12	59.53	59.03
	3	50.75	56.69	59.71	59.07
	4	50.50	56.41	59.43	58.60
WAIC	1	51.52	57.47	60.34	59.59
	2	51.67	57.41	60.76	60.42
	3	51.54	57.64	60.82	60.28
	4	51.26	57.67	60.52	60.05

Table 2: Information criteria from various models, calculated from four Markov chain Monte Carlo simulations using different starting values. For clarity we have removed a fixed constant of 2200 from each criterion.

Figure 2 illustrates a posterior predictive checking procedure, as described in Section 1.8. We specify the test statistics for $j = 1, 2$ to be the average January maxima on the jth margin, where the average is taken over the dataset period 1994-2013. The average January maxima at Port Fairy is 38.4 degrees Celsius, and the average January maxima at Melbourne Airport is 39.4 degrees Celsius. The posterior predictive check in Figure 2 compares these data values to values simulated from the posterior predictive distribution. The 8 000 Markov chain iterations are used to estimate the posterior predictive distributions, as discussed in Section 13.1.2. It can be seen from Figure 2 that the posterior predictive distribution of the average January maxima is consistent with the data at both sites. Similar conclusions are reached when using the largest January maxima as the test statistic.

The posterior predictive distributions derived from the model are plotted against time in Figure 1 for the Port Fairy weather station. Figure 1 shows both the seasonality and the small increasing trend. Posterior density estimates can be easily calculated for any function of the parameters. Figure 3 shows some examples of posterior density estimates calculated from this model. The left plot of Figure 3 shows the posterior distribution of the shape parameter at each site. The shape parameter is clearly larger at Melbourne Airport and has a greater variability than the shape parameter at Port Fairy.

The right plot of Figure 3 shows posterior distributions for conditional upper quantiles at Port Fairy (for January). The conditional upper quantile is defined here as the monthly temperature maximum at Port Fairy that is exceeded with probability 0.05, conditioning on the fact that the monthly maximum at Melbourne Airport exceeds some temperature

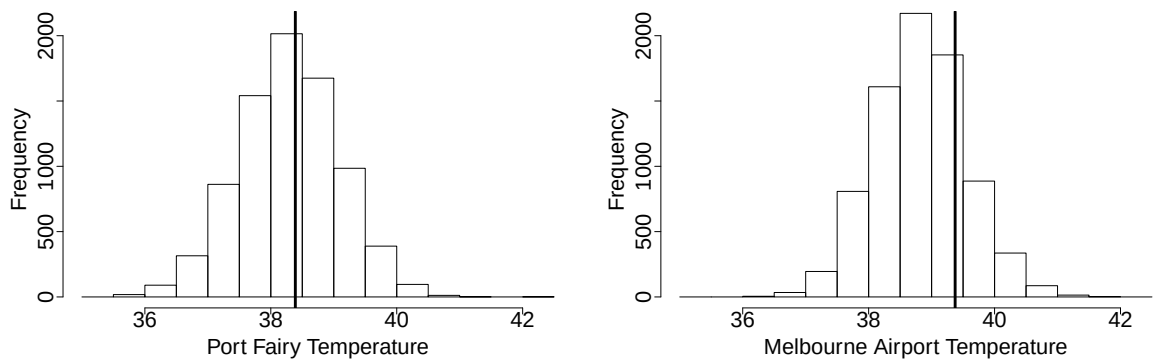

Figure 2: Posterior predictive distributions of the average January temperature maxima at Port Fairy (left) and Melbourne Airport (right), from the model described in this section. The solid vertical lines denote the values observed in the data.

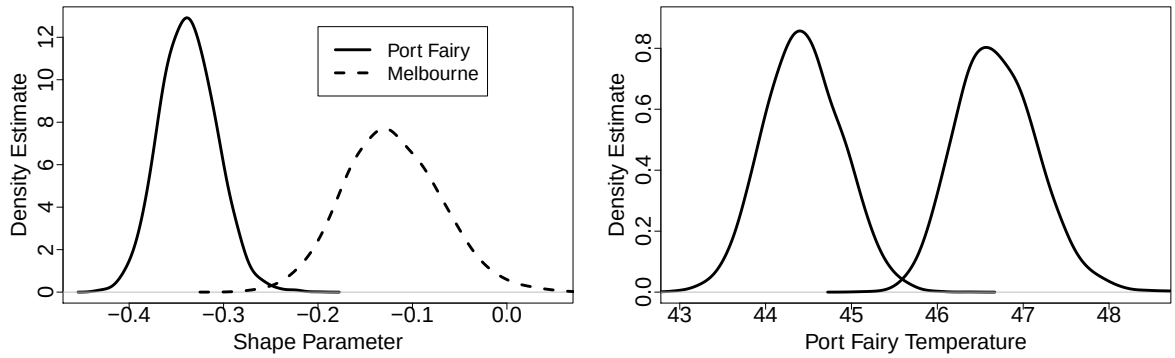

Figure 3: Left: Posterior distributions of shape parameters ξ_1 and ξ_2 at the Port Fairy and Melbourne sites respectively. Right: Posterior distributions (for January) for the monthly temperature maximum that is exceeded at Port Fairy with probability 0.05, conditioning on the fact that the monthly temperature maximum at Melbourne Airport exceeds 35 degrees Celsius (left curve) or 45 degrees Celsius (right curve).

x. The plot shows the distributions for $x = 35$ and $x = 45$. The conditional quantiles depend on both the dependence structure and the marginal distributions.

References

Bayarri, M. J. and Berger, J. O. (1999) Quantifying surprise in the data and model verification. In *Bayesian Statistics 6* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford University Press, pp. 475–501.

Bayarri, M. J. and Berger, J. O. (2000) P-values for composite null models. *J. Amer. Statist. Assoc.*, **95**, 1127–1142.

Boldi, M. O. and Davison, A. C. (2007) A mixture model for multivariate extremes. *J. R. Statist. Soc. B*, **69**, 217–229.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*. Boca Raton: CRC Press, 3rd edn.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. To Appear.

Gelman, A., Meng, X. L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733–807.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. B*, **64(4)**, 583–639.

Tawn, J. A. (1988) Bivariate extreme value theory: models and estimation. *Biometrika*, **75**, 397–415.

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Index