1 Distribution of the SNHT Statistic

Suppose we have a time series X_i where $X_i \sim N(\mu_i, \sigma^2)$ and

$$\mu_i = f_1(\text{Day of Year}) + f_2(\text{Hour}) + \beta(\text{Year})$$

From here on, I'll use d_i, h_i, and y_i to represent the day of year, hour of day, and year for the ith observation. Now, we'll make some additional assumptions about the data:

- We have two observations per day for a period of several years.
- The distribution of the hour for each observation is identical.

Also, the X_i may also be correlated in time. Now, consider the test statistic as defined by Haimberger (2007):

$$\begin{aligned}
T &= \frac{N}{2s_i}\left((\bar{X}_{Li} - \bar{X})^2 + (\bar{X}_{Ri} - \bar{X})^2\right) \\
&= \frac{N}{2s_i}\left((\bar{X}_{Li} - (\bar{X}_{Li} + \bar{X}_{Ri})/2)^2 + (\bar{X}_{Ri} - (\bar{X}_{Li} + \bar{X}_{Ri})/2)^2\right) \\
&= \frac{N}{2s_i}\left((\bar{X}_{Li} - \bar{X}_{Ri})^2/4 + (\bar{X}_{Ri} - \bar{X}_{Li})^2/4\right) \\
&= \frac{N(\bar{X}_{Li} - \bar{X}_{Ri})^2}{4s_i}
\end{aligned}$$

In order to have 1 year of observations, we set $N = 730$. $\bar{X}_{Li}$ is the mean of many normal random variables, and so $\bar{X}_{Li}$ is also normal. Moreover,

$$\begin{aligned}
E(\bar{X}_{Li}) &= \frac{1}{N}\sum_{j=1}^{N} E(X_{i-j}) \\
&= \frac{1}{N}\sum_{j=1}^{N} E(E(X_{i-j}|h_{i-j}, d_{i-j})) \\
&= \frac{1}{N}\sum_{j=1}^{N} E(f_1(d_{i-j}) + f_2(h_{i-j}) + \beta y_{i-j}) \\
&= \frac{1}{N}\sum_{j=1}^{N} f_1(d_{i-j}) + E(f_2(h_{i-j})) + \beta y_{i-j}
\end{aligned}$$

Likewise, we have

$$E(\bar{X}_{Ri}) = \frac{1}{N}\sum_{j=1}^{N} f_1(d_{i+j}) + E(f_2(h_{i+j})) + \beta y_{i+j}$$

and so

$$\begin{aligned}
E(\bar{X}_{Li} - \bar{X}_{Ri}) &= \frac{1}{N}\sum_{j=1}^{N} [f_1(d_{i-j}) + E(f_2(h_{i-j})) + \beta y_{i-j}] \\
&- \frac{1}{N}\sum_{j=1}^{N} [f_1(d_{i+j}) + E(f_2(h_{i+j})) + \beta y_{i+j}] \\
&= \frac{1}{N}\sum_{j=1}^{N} [f_1(d_{i+j-N-1}) - f_1(d_{i+j}) + E(f_2(h_{i+j-N-1}) - f_2(h_{i+j})) + \beta(y_{i+j-N-1} - y_{i+j} \\
&= \frac{1}{N}\sum_{j=1}^{N} \beta \\
&= \beta
\end{aligned}$$

where the hourly effects drop out because the hourly distributions are assumed the same year to year, the daily effects drop out because $d_{i+j-N-1} = d_{i+j}$, and we use the fact that $y_{i+j-N-1} = y_{i+j} - 1$.

Now, we compute the variance:

$$\begin{aligned} Var(\bar{X}_{Li} - \bar{X}_{Ri}) &= Var\left(\frac{1}{N}\sum_{j=1}^{N} X_{i-j} - \frac{1}{N}\sum_{j=1}^{N} X_{i+j}\right) \\ &= \frac{1}{N^2} Var\left(\sum_{j=1}^{N} X_{i-j} - \sum_{j=1}^{N} X_{i+j}\right) \\ &= \frac{(\mathbf{1}, \mathbf{-1})\Sigma_{(-i)}(\mathbf{1}, \mathbf{-1})^T}{N^2} \end{aligned}$$

where $\mathbf{1}$ is a vector of length N with every element 1, and $\Sigma_{(-i)}$ is the $2N \times 2N$ covariance matrix between the N observations before and the N observations following observation i.

Now, we make an assumption: the covariance between two observations depends only on the length of time between them. This allows us to compute $\Sigma_{(-i)}$ given a model for the autocorrelation function, $\hat{\phi}$. Thus, we finally conclude

$$\bar{X}_{Li} - \bar{X}_{Ri} \sim N(\beta, (\mathbf{1}, \mathbf{-1})\Sigma_{(-i)}(\mathbf{1}, \mathbf{-1})^T / N^2)$$

Thus, instead of Haimburger's original statistic, we will use

$$T = \frac{N^2(\bar{X}_{Li} - \bar{X}_{Ri})^2}{(\mathbf{1}, \mathbf{-1})\Sigma_{(-i)}(\mathbf{1}, \mathbf{-1})^T}$$

Since $\bar{X}_{Li} - \bar{X}_{Ri}$ is normal, we conclude

$$T \sim \chi_1^2$$

However, suppose we instead model the seasonality of the data, and that this model removes the seasonality and auto-correlation. In this scenario, each X_i can now be assumed $N(\mu, \sigma^2)$. In this case, we have

$$
\begin{aligned}
\bar{X}_{Li} &\sim N(\mu, \sigma^2/N) \\
\bar{X}_{Ri} &\sim N(\mu, \sigma^2/N) \\
Cov(\bar{X}_{Li}, \bar{X}_{Ri}) &= 0 \\
(\bar{X}_{Li} - \bar{X}_{Ri}) &\sim N(0, 2\sigma^2/N) \\
T &= \frac{N(\bar{X}_{Li} - \bar{X}_{Ri})^2}{2\sigma^2} \sim \chi^2_1
\end{aligned}
$$

2 Multiple Testing Correction

We compute the SNHT statistic for most observations in the dataset (all observations except those close to the edge). Thus, we have a multiple testing problem. Moreover, each test statistic is strongly positively correlated with many others, as they use most of the same data. Thus, we need to adjust the rejection threshold so as to ensure that we control the false discovery rate to at most α.

Suppose we compute N_s statistics for our dataset. One approach is to use the Benjamini-Hochberg correction. This correction works by sorting the p-values and then rejecting the i smallest p-values such that $p_{(i)} \leq \alpha/(N_s - i + 1)$.

Let's see if we can derive the exact distribution of the maximum of the test statistics. Under H_0, we have

$$T_i = \frac{N \cdot (\bar{X}_{Li} - \bar{X}_{Ri})}{\sqrt{(\mathbf{1}, -\mathbf{1})\Sigma_{(-i)}(\mathbf{1}, -\mathbf{1})^T}} \sim N(0, 1)$$

So, we have N_s standard normal random variates, but they are not independent. However, we first assume they are independent. Let $Y = \max(T_i)$, then

$$
\begin{aligned}
P(Y \leq k) &= P(T_1 \leq k)P(T_2 \leq k) \cdots P(T_{N_s} \leq k) \\
&= P(T_1 \leq k)^{N_s} \\
f_Y(y) &= N_s F_T(t)^{N_s - 1} f_T(t)
\end{aligned}
$$

Now, accounting for independence:

$$
\begin{aligned}
\operatorname{Cov}(T_i, T_j) &= \operatorname{Cov}\left(\frac{N \cdot (\bar{X}_{Li} - \bar{X}_{Ri})}{\sqrt{(\mathbf{1}, \mathbf{-1})\Sigma_{(-i)}(\mathbf{1}, \mathbf{-1})^T}}, \frac{N \cdot (\bar{X}_{Lj} - \bar{X}_{Rj})}{\sqrt{(\mathbf{1}, \mathbf{-1})\Sigma_{(-j)}(\mathbf{1}, \mathbf{-1})^T}}\right) \\
&= \frac{N^2}{(\mathbf{1}, \mathbf{-1})\Sigma_{(-i)}(\mathbf{1}, \mathbf{-1})^T}\operatorname{Cov}\left(\bar{X}_{Li} - \bar{X}_{Ri}, \bar{X}_{Lj} - \bar{X}_{Rj}\right)
\end{aligned}
$$

as $\Sigma_{(-i)}$ and $\Sigma_{(-j)}$ are determined via the model for the auto-correlation function. Now, $\bar{X}_{Li}$ and $\bar{X}_{Lj}$ have $k = \max(N - |j - i|, 0)$ overlapping terms. So

$$
\begin{aligned}
\operatorname{Cov}(\bar{X}_{Li}, \bar{X}_{Lj}) &= \operatorname{Cov}\left(\frac{1}{N}\sum_{t_i=1}^{N} X_{t_i}, \frac{1}{N}\sum_{t_j=1}^{N} X_{t_j}\right) \\
&= \frac{1}{N^2}\operatorname{Cov}\left(\sum_{t_i=1}^{N} X_{t_i}, \sum_{t_j=1}^{N} X_{t_j}\right)
\end{aligned}
$$