STAT 153 & 248 - Time Series
Lecture Fourteen

Spring 2025, UC Berkeley

Aditya Guntuboyina

March 6, 2025

We shall discuss three simple high-dimensional models for time series $y_1, \dots, y_n$. The first model was already discussed last week, the second model is new but very simple, and the third model is the same as the spectrum model from last lecture. The second model makes it easier to understand the third model.

1 Model One

This is the model

$$y_t \overset{\text{ind}}{\sim} N(\mu_t, \sigma^2)$$

where ind stands for "independently distributed as". Note that the right hand side depends on t so the distribution of y_t changes with t and we cannot therefore use "i.i.d".

The parameters in this model are $\mu_1, \dots, \mu_n$ and σ^2. Clearly this is a high-dimensional because the number of parameters is large.

If we attempt to estimate the parameters by maximizing the likelihood without any regularization, we get $\mu_t = y_t$ and $\sigma^2 = 0$, leading to full interpolation (overfitting) to the data. Regularization is therefore necessary to obtain something useful. If we want to obtain "smooth" trend estimates, we can employ regularization terms which force neighboring values or neighboring slopes of μ_t to be close. If we focus on slopes (which leads to more smoothness compared to just imposing closeness of values), we obtain the estimators $\hat{\mu}_t^{\text{ridge}}(\lambda)$ and $\hat{\mu}_t^{\text{lasso}}(\lambda)$ which minimize:

$$\sum_{t=1}^{n} (y_t - \mu_t)^2 + \lambda \sum_{t=2}^{n-1} \left((\mu_{t+1} - \mu_t) - (\mu_t - \mu_{t-1}) \right)^2$$

and

$$\sum_{t=1}^{n} (y_t - \mu_t)^2 + \lambda \sum_{t=2}^{n-1} \left| (\mu_{t+1} - \mu_t) - (\mu_t - \mu_{t-1}) \right|$$

respectively. We have already studied these estimators last week where we observed, among other things, that they can be alternatively represented as $\hat{\mu}_t^{\text{ridge}}(\lambda) = X \hat{\beta}^{\text{ridge}}(\lambda)$ and

$\hat{\mu}_t^{\text{lasso}}(\lambda) = X\hat{\beta}^{\text{lasso}}(\lambda)$ where

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 2 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n-1 & n-2 & \cdot & \cdot & \cdot & 1 \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{n-1} \end{pmatrix} \tag{1}$$

and $\hat{\beta}^{\text{ridge}}(\lambda)$ and $\hat{\beta}^{\text{lasso}}(\lambda)$ minimize

$$\|y - X\beta\|^2 + \sum_{t=2}^{n-1} \beta_t^2$$

and

$$\|y - X\beta\|^2 + \sum_{t=2}^{n-1} |\beta_t|$$

respectively.

This model is an example of a “Mean Model” where the focus is on estimating the mean parameters μ_t. In contrast, the next two models will be examples of “Variance Models”.

2 Model Two

This is the model

$$y_t \overset{\text{ind}}{\sim} N(0, \tau_t^2) \tag{2}$$

The parameters are $\tau_1^2, \dots, \tau_n^2$. Since these represent variances, we refer to this as a “variance model”. τ_t can be interpreted as the magnitude of y_t. This model is useful when we care only about the magnitudes of the observations y_t (and not their signs).

Model (2) (similar to Model One from the previous section) is also a high-dimensional model because the number of parameters is large.

The likelihood is proportional to

$$\prod_{t=1}^{n} \frac{1}{\tau_t} \exp\left(-\frac{y_t^2}{2\tau_t^2}\right). \tag{3}$$

Note that the likelihood depends on the data only through the squares $y_1^2, \dots, y_n^2$. Therefore the squares $y_1^2, \dots, y_n^2$ form the “sufficient statistic” in this model. Under (2), we have

$$y_t^2 \overset{\text{ind}}{\sim} \tau_t^2 \chi_1^2$$

where χ_1^2 denotes the chi-squared distribution with 1 degree of freedom. Instead of writing the likelihood in terms of the raw data y_t, we can also write the likelihood using the squares y_t^2. This will lead to a slightly different form for the likelihood that should still be proportional to (3). To see this, observe that the density of χ_1^2 is proportional to $x^{-1/2}\exp(-x/2)$ so that the density of $\tau_t^2\chi_1^2$ is proportional to

$$\frac{1}{\tau_t^2}\left(\frac{x}{\tau_t^2}\right)^{-1/2} \exp\left(-\frac{x}{2\tau_t^2}\right) = x^{-1/2}\frac{1}{\tau_t}\exp\left(-\frac{x}{2\tau_t^2}\right).$$

The likelihood written in terms of $y_1^2, \ldots, y_n^2$ is thus

$$\prod_{t=1}^n (y_t^2)^{-1/2} \frac{1}{\tau_t} \exp\left(-\frac{y_t^2}{2\tau_t^2}\right).$$

The term $(y_t^2)^{-1/2}$ above can be dropped as it is a constant not depending on the parameters τ_t^2. Dropping it leads to (3).

The log-likelihood is:

$$\sum_{t=1}^n \left(-\log \tau_t - \frac{y_t^2}{2\tau_t^2}\right).$$

It is a convention to write optimization problems for computing estimators as minimization problems (as opposed to maximization). For this, we write the negative log-likelihood which is given by:

$$\sum_{t=1}^n \left(\log \tau_t + \frac{y_t^2}{2\tau_t^2}\right).$$

As τ_t is a standard deviation parameter that is constrained to be positive, it is better to deal with $\alpha_t = \log \tau_t$ instead of τ_t directly. This reparameterization has the following benefits:

- **Unconstrained Optimization**: Unlike τ_t, which must be positive, α_t can take any real value, allowing for more stable numerical optimization.
- **Improved Computational Stability**: Variance parameters can vary over several orders of magnitude, and working in the log scale reduces numerical precision issues.

Because of these benefits, many variance modeling approaches use log-variance transformations (see e.g., stochastic volatility models).

Writing the negative log-likelihood in terms of $\alpha_t = \log \tau_t$, we get

$$\sum_{t=1}^n \left(\alpha_t + \frac{y_t^2}{2} e^{-2\alpha_t}\right)$$

If we minimize the above (without any additional regularization) with respect to α_t, we obtain $\alpha_t = \log |y_t|$, or equivalently, $\tau_t^2 = y_t^2$. In other words, the parameters τ_t^2 will fully interpolate (overfit) the sufficient statistics y_t^2.

For a more useful estimation procedure, we need to introduce regularization. If we assume that α_t is smooth, we can add the penalty $\sum_{t=2}^{n-1} ((\alpha_{t+1} - \alpha_t) - (\alpha_t - \alpha_{t-1}))^2$ or $\sum_{t=2}^{n-1} |(\alpha_{t+1} - \alpha_t) - (\alpha_t - \alpha_{t-1})|$ to the negative log-likelihood. This leads to the estimators $\hat{\alpha}_t^{\text{ridge}}(\lambda)$ and $\hat{\alpha}_t^{\text{lasso}}(\lambda)$ which are defined as the minimizers of

$$\sum_{t=1}^n \left(\alpha_t + \frac{y_t^2}{2} e^{-2\alpha_t}\right) + \lambda \sum_{t=2}^{n-1} ((\alpha_{t+1} - \alpha_t) - (\alpha_t - \alpha_{t-1}))^2$$

and

$$\sum_{t=1}^n \left(\alpha_t + \frac{y_t^2}{2} e^{-2\alpha_t}\right) + \lambda \sum_{t=2}^{n-1} |(\alpha_{t+1} - \alpha_t) - (\alpha_t - \alpha_{t-1})|$$

respectively. The penalties encourage smoothness in $\{\alpha_t\}$, leading to more stable and interpretable variance estimates. These optimizations are convex and they can be solved, for example, using functions from the python library `cvxpy` just as we solved $\hat{\beta}^{\text{ridge}}(\lambda)$ and $\hat{\beta}^{\text{lasso}}(\lambda)$ from the previous section.

3 Model Three

Model three is essentially Model two but applied to the DFT. In order to describe this, let us first revisit the DFT. Given a time series $y_0, \dots, y_{n-1}$, its DFT is $b_0, b_1, \dots, b_{n-1}$ where

$$b_j = \sum_{t=0}^{n-1} y_t \exp\left(-\frac{2\pi i j t}{n}\right).$$

b_j is a complex number with real and imaginary parts given by

$$\mathrm{Re}(b_j) = \sum_{t=0}^{n-1} y_t \cos\left(\frac{2\pi j t}{n}\right) \quad \text{and} \quad \mathrm{Im}(b_j) = -\sum_{t=0}^{n-1} y_t \sin\left(\frac{2\pi j t}{n}\right)$$

When $j = 0$, the imaginary part is zero and we get $b_0 = \sum_{t=0}^{n-1} y_t$. b_0 is therefore just the sum of the datapoints and it does not provide any information on cycles etc.

Another fact that we previously verified is $b_{n-j} = \bar{b_j}$ (here $\bar{b_j}$ denotes the complex conjugate of b_j). Because of this property, the later half of the DFT terms is redundant (as they can be recovered from the first half).

If n is odd and $m = (n-1)/2$, then the most important DFT terms are $b_1, \dots, b_m$. The other terms are b_0 which is simply the sum of the data points and $b_{m+1}, \dots, b_{n-1}$ which are simply the complex conjugates of $b_m, b_{m-1}, \dots, b_1$.

If n is even and $m = (n-2)/2$, then the most important DFT terms are $b_1, \dots, b_m$ and b_{m+1}. The other DFT terms are b_0 which is simply the sum of the data points and $b_{m+2}, \dots, b_{n-1}$ which are the complex conjugates of $b_m, \dots, b_1$. Note in this case that ${}_{m+1} = b_{n/2}$ will have zero imaginary part (hence b_{m+1} is real).

Below we focus on the case where n is odd for simplicity, and take $m = (n-1)/2$. Model three is obtained by using Model two for the DFT terms $b_1, \dots, b_m$. Because b_j can be complex, we use the modeling assumption for both the real and imaginary parts:

$$\mathrm{Re}(b_j), \mathrm{Im}(b_j) \overset{\text{i.i.d}}{\sim} N(0, \gamma_j^2) \qquad \text{for } j = 1, \dots, m. \tag{4}$$

We also assume that b_j are independent across j. The unknown parameters in this model are $\gamma_1, \dots, \gamma_m$. γ_j represents the strength of the sinusoids at frequency j/n.

The likelihood corresponding to (4) is proportional to:

$$\prod_{j=1}^{m} \frac{1}{\gamma_j} \exp\left(-\frac{(\mathrm{Re}(b_j))^2}{2\gamma_j^2}\right) \frac{1}{\gamma_j} \exp\left(-\frac{(\mathrm{Im}(b_j))^2}{2\gamma_j^2}\right)$$
$$= \prod_{j=1}^{m} \frac{1}{\gamma_j^2} \exp\left(-\frac{(\mathrm{Re}(b_j))^2 + (\mathrm{Im}(b_j))^2}{2\gamma_j^2}\right) = \prod_{j=1}^{m} \frac{1}{\gamma_j^2} \exp\left(-\frac{|b_j|^2}{2\gamma_j^2}\right).$$

Therefore the likelihood depends on the squared magnitudes $|b_j|^2$ of the DFT coefficients. Recall that the periodogram $I(j/n)$ is defined as

$$I(j/n) := \frac{|b_j|^2}{n}.$$

We can therefore rewrite the likelihood in terms of the periodogram as follows:

$$\prod_{j=1}^{m} \frac{1}{\gamma_j^2} \exp\left(-\frac{nI(j/n)}{2\gamma_j^2}\right). \tag{5}$$

This likelihood depends on the data only through the periodogram ordinates $I(j/n)$ for $j = 1, \dots, m$. Therefore the periodogram forms the sufficient statistic in this model. Under (4), we have

$$I(j/n) = \frac{1}{n}|b_j|^2 = \frac{1}{n}\left((\mathrm{Re}(b_j))^2 + (\mathrm{Im}(b_j))^2\right) \sim \frac{\gamma_j^2}{n}\chi_2^2.$$

The model can therefore be written directly in terms of the periodogram as

$$I(j/n) \overset{\text{ind}}{\sim} \frac{\gamma_j^2}{n}\chi_2^2 \qquad \text{for } j = 1, \dots, m.$$

We can write the likelihood for the above model in terms of the periodogram and this would be proportional to (5). Note also that χ_2^2 distribution with two degrees of freedom actually coincides with the Exponential distribution with λ parameter equal to 1/2.

Intuitively, Model (4) does not care so much about the individual DFT coefficients b_j but only their magnitude.

The negative log-likelihood corresponding to (5) is

$$\sum_{j=1}^{m}\left(2\log\gamma_j + \frac{nI(j/n)}{2\gamma_j^2}\right).$$

As in the case of Model two, for optimization purposes we work with the logarithms of γ_j. Let $\alpha_j = \log\gamma_j$. The negative log-likelihood in terms of α_j is

$$\sum_{j=1}^{m}\left(2\alpha_j + \frac{nI(j/n)}{2}e^{-2\alpha_j}\right).$$

If we directly minimize the above with respect to α_j (without any additional regularization), we get

$$\alpha_j = \log\sqrt{\frac{nI(j/n)}{2}} \qquad \text{and} \qquad \gamma_j^2 = e^{2\alpha_j} = \frac{nI(j/n)}{2}.$$

This basically means that the γ_j^2 parameters fully interpolate the periodogram leading to full overfitting. For more meaningful estimation, we need to add regularization. If we assume smoothness of α_j, we can add the penalty $\sum_{j=2}^{m-1}((\alpha_{j+1} - \alpha_j) - (\alpha_j - \alpha_{j-1}))^2$ or $\sum_{j=2}^{m-1}|(\alpha_{j+1} - \alpha_j) - (\alpha_j - \alpha_{j-1})|$ to the negative log-likelihood. This leads to the estimators $\hat{\alpha}_t^{\text{ridge}}(\lambda)$ and $\hat{\alpha}_t^{\text{lasso}}(\lambda)$ which are defined as the minimizers of

$$\sum_{j=1}^{m}\left(2\alpha_j + \frac{nI(j/n)}{2}e^{-2\alpha_j}\right) + \lambda\sum_{j=2}^{m-1}((\alpha_{j+1} - \alpha_j) - (\alpha_j - \alpha_{j-1}))^2$$

and

$$\sum_{t=1}^{n}\left(2\alpha_j + \frac{nI(j/n)}{2}e^{-2\alpha_j}\right) + \lambda\sum_{j=2}^{m-1}|(\alpha_{j+1} - \alpha_j) - (\alpha_j - \alpha_{j-1})|$$

respectively. The penalties encourage smoothness in $\{\alpha_j\}$, leading to more stable and interpretable estimates for $\{\gamma_j\}$.

In the next lecture, we shall explain the equivalence of this model with the spectrum model from last lecture. We shall also explore some applications for this model.