Imputation of Time Series with Missing Values under Heavy-Tailed AR Model via Stochastic EM

Prof. Daniel P. Palomar
The Hong Kong University of Science and Technology (HKUST)

Invited Talk - International Workshop on
Mathematical Issues in Information Sciences (MIIS)
The Chinese University of Hong Kong, Shenzhen
21 December 2018

Outline

1. **Motivation**
2. **Imputation: State of the art**
   - Different formulations
   - Basics on imputation
3. **Approach for statistical time series imputation**
   - Step 1: Estimation of parameters
   - Step 2: Imputation of missing values
4. **Numerical simulations**
5. **Summary**

Outline

Data with missing values

- In theory, data is typically assumed complete and algorithms are designed for complete data.
- In practice, however, **often data has missing values**, due to a variety of reasons.
- Then the algorithms designed for complete data **can be disastrous!**
- Missing values typically happen during the data observation or recording process:[1]
  - values may not be measured,
  - values may be measured but get lost, or
  - values may be measured but are considered unusable.

[1] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd Ed. Hoboken, N.J.: John Wiley & Sons, 2002.

Data with missing values

- Some real-world cases where missing values occur:
  - some stocks may suffer a lack of liquidity resulting in no transaction and hence no price recorded
  - observation devices like sensors may break down during the measurement
  - weather or other conditions may disturb sample taking schemes
  - in industrial experiments some results may be missing because of mechanical breakdowns unrelated to the experimental process
  - in an opinion survey some individuals may be unable to express a preference for one candidate over another
  - respondents in a survey may not answer every question
  - countries may not collect statistics every year
  - archives may simply be incomplete
  - subjects may drop out of panels.

What is imputation?

- How can we cope with data with missing values?
- One option is to design processing algorithms that can accept missing values, but has to be done in a case by case basis and is expensive.
- Another option is **imputation**: filling in those missing values based on some properties of the data. After that, processing algorithms for complete data can be safely used.
- However, magic cannot be done to impute missing values. One has to rely on some structural properties like some temporal structure.
- There are many imputation techniques, many heuristic (can do more harm than good) and some with a sound statistical foundation.
- Many works assume a Gaussian distribution, which doesn't hold in many applications.

👉 *We will focus on statistically sound methods for time series imputation under heavy-tailed distributions.*

Outline

Outline

Netflix problem: Low-rank matrix completion

- In big data system analytics, it is often the case that the high-dimensional data matrix lies in a low-dimensional subspace.
- A popular example is the Netflix problem where the data matrix contains movie ratings by users and is extremely sparse:

$$\mathbf{X} = \overbrace{\begin{bmatrix} 2 & 3 & ? & ? & 5 & ? \\ 1 & ? & ? & 4 & ? & 3 \\ ? & ? & 3 & 2 & ? & 5 \\ 4 & ? & 3 & ? & 2 & 4 \end{bmatrix}}^{\text{Movies}} \text{Users}$$

- In 2009, the Netflix prize of US$1M was awarded to a method based, among other techniques, on the **low-rank property**:

$$\mathbf{X} = \mathbf{A}\mathbf{B}^T$$

  where both **A** and **B** are extremely thin matrices.
- This low-rank property can be used to impute missing values.

Inpainting in image processing

- In image processing, a popular problem is that of images with missing blocks of pixels:

- In this case, one can use the natural structure of images, e.g., small gradient or a dictionary of small structures commonly appearing.
- **Total variation** is a common technique that imputes the missing pixels by ensuring a small ℓ_1-norm of the gradient.
- Learning an **overcomplete dictionary** allows for imputing blocks of pixels based on the dictionary.

Frugal sensing and compressive covariance sensing

- A somehow related problem in signal processing and wireless communications is frugal sensing and compressive covariance sensing where one wants to obtain the complete knowledge of a covariance matrix.
- In frugal sensing,[2] one wants to obtain the matrix $\mathbf{X}$ from knowledge of the value of some cuts $\mathbf{c}_i^T \mathbf{X} \mathbf{c}_i = v_i$ or even just one bit of information of the cuts $\mathbf{c}_i^T \mathbf{X} \mathbf{c}_i \lesseqgtr t$.
- More generally, in compressive covariance sensing,[3] one wants to reconstruct $\mathbf{X}$ from the smaller matrix $\mathbf{C}^T \mathbf{X} \mathbf{C}$, where $\mathbf{C}$ is some tall compression or selection matrix that exploits structural information or sparsity in some domain.

---

[2] O. Mehanna and N. Sidiropoulos, "Frugal sensing: Wideband power spectrum sensing from few bits," *IEEE Trans. on Signal Processing*, vol. 61, no. 10, pp. 2693–2703, 2013.

[3] D. Romero, D. D. Ariananda, Z. Tian, and G. Leus, "Compressive covariance sensing: Structure-based compressive sensing beyond sparsity," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 78–93, 2016.

Time series with structure

- In some applications, one of the dimensions of the data matrix is time.
- The time dimension sometimes has some specific structure on the distribution of the missing values, like the monotone missing pattern.[4]
- The time dimension can also follow some structural model that can be effectively used to fill in the missing values.
- One simple example of **time structure** is the **random walk**, which is pervasive in financial applications (e.g., log-returns of stocks):[5]

$$y_t = \phi_0 + y_{t-1} + \varepsilon_t.$$

- Another example of **time structure** is the **AR(p) model** (e.g., traded log-volume of stocks):

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t.$$

[4] J. Liu and D. P. Palomar, "Robust estimation of mean and covariance matrix for incomplete data in financial applications," in *Proc. IEEE GlobalSIP*, Montreal, Canada, 2017.

[5] J. Liu, S. Kumar, and D. P. Palomar, "Parameter estimation of heavy-tailed random walk model from incomplete data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

Outline

Simple imputation methods

Some simple and naive approaches for imputation are:

- **hot deck imputation**: recorded units in the sample are used to substitute values
- **mean imputation**: means from sets of recorded values are substituted

👉 *sounds like a good idea but it distorts the empirical distribution of the sampled values (biases in variances and covariances)* 😱

$$y_1, y_2, \dots, y_{k1}, \mathrm{NA}_1, \dots, \mathrm{NA}_{k2} \rightarrow \hat{\sigma}^2 = \frac{1}{k1}\sum_{i=1}^{k1}(y_i - \hat{\mu})^2$$

$$y_1, y_2, \dots, y_{k1}, \hat{\mu}, \dots, \hat{\mu} \rightarrow \hat{\sigma}^2 = \frac{1}{k1+k2}\left(\sum_{i=1}^{k1}(y_i - \hat{\mu})^2 + \sum_{i=1}^{k2} 0\right)$$

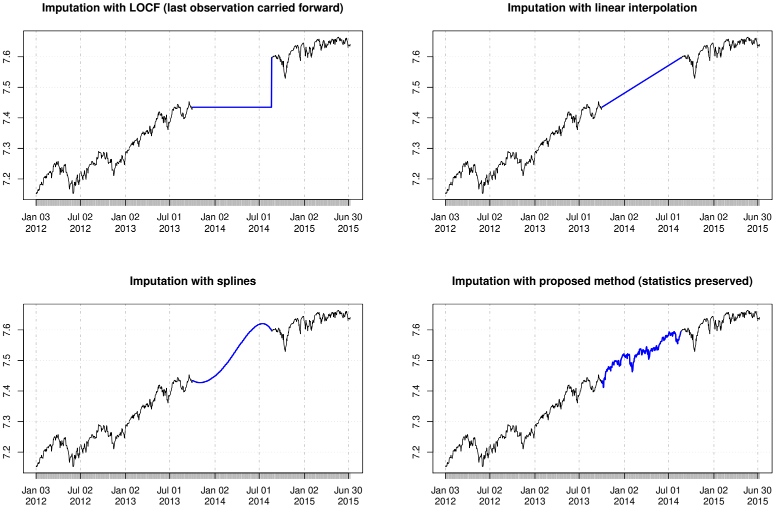
Naive vs sound imputation methods

- Ad-hoc methods of imputation can lead to serious biases in variances and covariances.
- Examples are:
  - mean imputation
  - constant interpolation
  - linear interpolation
  - polynomial interpolation
  - spline interpolation
- A sound imputation method should preserve the statistics of the observed values.

👉 *The statistical way is to first estimate the distribution of the missing values conditional on the observed values* $f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}})$ *and then impute based on that posterior distribution.*

Naive vs sound time series imputation methods

Illustration of different naive imputation methods and a sound statistical method that preserves the statistics:



Single and multiple imputation

- Suppose we have somehow estimated the conditional distribution $f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}})$.
- At this point it is trivial to randomly generate the missing values from that distribution:

$$\mathbf{y}_{\text{miss}} \sim f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}}).$$

- This only gives you one realization of the missing values.
- In some applications, one would like to have multiple realizations of the missing values to properly test the performance of some subsequent methods or algorithms.
- Multiple imputation (MI) consists of generating multiple realizations of the missing values:

$$\mathbf{y}_{\text{miss}}^{(k)} \sim f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}}) \qquad \forall k = 1, \ldots, K.$$

Outline

Outline

Estimation of parameters for iid Gaussian

- Suppose a univariate random variable y follows a Gaussian distribution:

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right).$$

- We have T incomplete samples $\{y_t\}$ and the missing mechanism is ignorable (aka MAR), the ML estimation problem is formulated as

$$\underset{\mu,\sigma^2}{\text{maximize}} \quad \log\left(\prod_{t \in C_{\text{obs}}} f_G\left(y_t; \mu, \sigma^2\right)\right),$$

where C_{obs} is the set of the indexes of the observed samples, and $f_G(\cdot)$ is the pdf of the Gaussian distribution.

- **Closed-form solution**: 😃

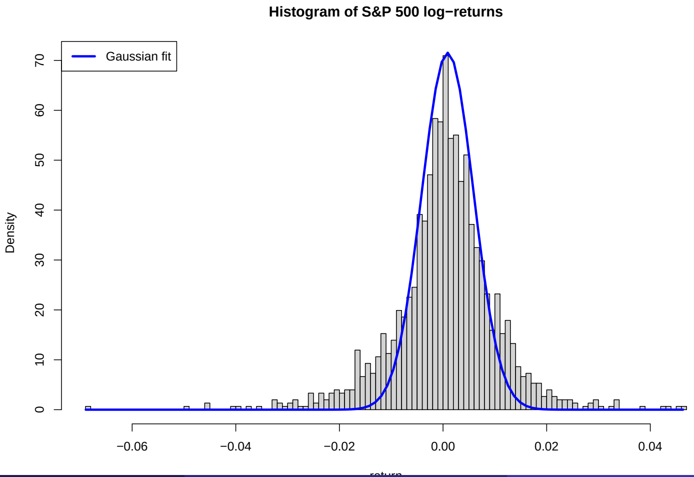
$$\hat{\mu} = \frac{1}{n_{obs}} \sum_{t \in C_{\text{obs}}} y_t$$

and

$$\hat{\sigma}^2 = \frac{1}{n_{\text{obs}}} \sum_{t \in C_{\text{obs}}} \left(y_t - \hat{\mu}\right)^2.$$

Estimation of parameters for iid Student's t

- In many applications, the **Gaussian distribution** is **not appropriate** and a more realistic heavy-tailed distribution is necessary.
- An example is in the financial returns of stocks:



Estimation of parameters for iid Student's t

- The Student's t-distribution is a widely used heavy-tailed distribution.
- Suppose y follows a Student's t-distribution: $y \sim t\left(\mu, \sigma^2, \nu\right)$ with pdf

$$f_t\left(y; \mu, \sigma^2, \nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right).$$

- Given the incomplete data set, the ML estimation problem for $\boldsymbol{\theta} = \left(\mu, \sigma^2, \nu\right)$ can be formulated as

$$\underset{\mu,\sigma^2,\nu}{\text{maximize}} \quad \log\left(\prod_{t\in C_{obs}} f_t\left(y_t; \mu, \sigma^2, \nu\right)\right).$$

  No closed-form solution. 😞

- Interestingly, the Student's t-distribution can be represented as a Gaussian mixture:

$$y_t|\tau_t \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\tau_t}\right), \quad \tau_t \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right),$$

  where τ_t is the mixture weight.

Estimation of parameters for iid Student's t

We can use the expectation-maximization (EM) algorithm to solve this ML estimation problem by regarding $\boldsymbol{\tau}_{\text{obs}} = \{\tau_t\}_{t \in C_{\text{obs}}}$ as latent variables:

- **Expectation (E) - step**: compute the expected complete data log-likelihood given the current estimates

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) =& \mathsf{E}_{f\left(\boldsymbol{\tau}_{\text{obs}}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\right)}\left[\log\left(f\left(\mathbf{y}_{\text{obs}}, \boldsymbol{\tau}_{\text{obs}} \mid \boldsymbol{\theta}\right)\right)\right] \\
=& -\sum_{t \in C_{\text{obs}}} \frac{w_t^{(k)}}{2\sigma^2}\left(y_t-\mu\right)^2 - \frac{n_{\text{obs}}}{2}\log\left(\sigma^2\right) + n_{\text{obs}}\frac{\nu}{2}\log\left(\frac{\nu}{2}\right) \\
& - n_{\text{obs}}\log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{\nu}{2}\sum_{t \in C_{\text{obs}}}\left(\delta_t^{(k)} - w_t^{(k)}\right) + const.,
\end{aligned}
$$

where $w_t^{(k)} = \mathsf{E}\left[\tau_t\right] = \frac{\nu^{(k)}+1}{\nu^{(k)}+\left(y_t-\mu^{(k)}\right)^2/\left(\sigma^{(k)}\right)^2}$,

$\delta_t^{(k)} = \mathsf{E}\left[\log\left(\tau_t\right)\right] = \psi\left(\frac{\nu^{(k)}+1}{2}\right) - \log\left(\frac{\nu^{(k)}+\left(y_t-\mu^{(k)}\right)^2/\left(\sigma^{(k)}\right)^2}{2}\right).$

Estimation of parameters for iid Student's t

- **Maximization (M) - step**: update the estimates as

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$$

and has **closed-form solution**: 😃

$$\mu^{(k+1)} = \frac{\sum_{t \in C_{\text{obs}}} w_t^{(k)} y_t}{\sum_{t \in C_{\text{obs}}} w_t^{(k)}},$$

$$\left(\sigma^{(k+1)}\right)^2 = \frac{\sum_{t \in C_{\text{obs}}} w_t^{(k)} \left(y_t - \mu^{(k+1)}\right)^2}{n_{\text{obs}}},$$

$$\nu^{(k+1)} = \underset{\nu > 0}{\operatorname{argmax}}\ n_{\text{obs}} \left(\frac{\nu}{2} \log\left(\frac{\nu}{2}\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right) + \frac{\nu}{2} \sum_{t \in C_{\text{obs}}} \left(\delta_t^{(k)} - w_t^{(k)}\right).$$

Algorithm

Stochastic EM algorithm for iid Student's t:

Initialize $\mu^{(0)}$, $\left(\sigma^{(0)}\right)^2$, and $\nu^{(0)}$. Set $k = 0$.

repeat

$$w_t^{(k)} = \frac{\nu^{(k)} + 1}{\nu^{(k)} + \left(y_t - \mu^{(k)}\right)^2 / \left(\sigma^{(k)}\right)^2},$$

$$\mu^{(k+1)} = \frac{\sum_{t \in C_{\text{obs}}} w_t^{(k)} y_t}{\sum_{t \in C_{\text{obs}}} w_t^{(k)}},$$

$$\left(\sigma^{(k+1)}\right)^2 = \frac{\sum_{t \in C_{\text{obs}}} w_t^{(k)} \left(y_t - \mu^{(k+1)}\right)^2}{n_{\text{obs}}},$$

$$\nu^{(k+1)} = \underset{\nu > 0}{\operatorname{argmax}}\ Q\left(\mu^{(k+1)}, \left(\sigma^{(k+1)}\right)^2, \nu | \mu^k, \left(\sigma^k\right)^2, \nu^k\right)$$

$k \leftarrow k + 1$

until convergence

Estimation of parameters for AR(1) Student's t

- Consider an AR(1) time series with innovations following a Student's t-distribution:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim t\left(0, \sigma^2, \nu\right)$.

- The ML estimation problem for $\boldsymbol{\theta} = \left(\varphi_0, \varphi_1, \sigma^2, \nu\right)$ is formulated as

$$\underset{\phi_0, \varphi_1, \sigma^2, \nu}{\text{maximize}} \quad \log\left(\int \prod_{t=2}^{T} f_t\left(y_t; \varphi_0 + \varphi_1 y_{t-1}, \sigma^2, \nu\right) d\mathbf{y}_{\text{miss}}\right)$$

- The objective function involves an integral, and we have **no closed-form expression**. 😱
- As before, we can represent ε_t as

$$\varepsilon_t | \tau_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{\tau_t}\right), \quad \tau_t \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right),$$

and use the EM type algorithm to solve this optimization problem by regarding $\mathbf{y}_{\text{miss}}$ and $\boldsymbol{\tau} = \{\tau_t\}_{t=1,\dots,T}$ as the latent variables.

Estimation of parameters for AR(1) Student's t

- The complete data log-likelihood $f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}}, \boldsymbol{\tau} \mid \boldsymbol{\theta})$ belongs to the exponential family:

$$\begin{aligned} & f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}}, \boldsymbol{\tau} \mid \boldsymbol{\theta}) \\ = & h(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}}, \boldsymbol{\tau}) \exp\left(-\psi(\boldsymbol{\theta}) + \langle \mathbf{s}(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}}, \boldsymbol{\tau}), \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle\right) \end{aligned}$$

where

$$h(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}}, \boldsymbol{\tau}) = \prod_{t=2}^{T} \tau_t^{-\frac{1}{2}},$$

$$\psi(\boldsymbol{\theta}) = -(T-1)\left\{\frac{\nu}{2}\log\left(\frac{\nu}{2}\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{1}{2}\log\left(\sigma^2\right) - \frac{1}{2}\log(2\pi)\right\}$$

$$\mathbf{s}(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{miss}}, \boldsymbol{\tau}) = \sum_{t=2}^{T}\left[\log(\tau_t) - \tau_t, \tau_t y_t^2, \tau_t, \tau_t y_{t-1}^2, \tau_t y_t, \tau_t y_t y_{t-1}, \tau_t y_{t-1}\right],$$

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = \left[\frac{\nu}{2}, -\frac{1}{2\sigma^2}, -\frac{\varphi_0^2}{2\sigma^2}, -\frac{\varphi_1^2}{2\sigma^2}, \frac{\varphi_0}{\sigma^2}, \frac{\varphi_1}{\sigma^2}, -\frac{\varphi_0\varphi_1}{\sigma^2}\right].$$

Estimation of parameters for AR(1) Student's t

- Thus the expected complete data log-likelihood can be expressed as

$$\begin{aligned} Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right) =& \mathrm{E}_{f\left(\mathbf{y}_{\text{miss}},\boldsymbol{\tau}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\right)}\left[\log\left(f(\mathbf{y}_{\text{obs}},\mathbf{y}_{\text{miss}},\boldsymbol{\tau}\mid\boldsymbol{\theta})\right)\right] \\ =& -\psi\left(\boldsymbol{\theta}\right)+\left\langle \bar{\mathbf{s}}\left(\boldsymbol{\theta}^{(k)}\right),\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\right\rangle + const., \end{aligned}$$

where $\bar{\mathbf{s}}\left(\boldsymbol{\theta}^{(k)}\right) = \mathrm{E}_{f\left(\mathbf{y}_{\text{miss}},\boldsymbol{\tau}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\right)}\left[\mathbf{s}\left(\mathbf{y}_{\text{obs}},\mathbf{y}_{\text{miss}},\boldsymbol{\tau}\right)\right]$.

- The computation of $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$ is reduced to that of $\bar{\mathbf{s}}\left(\boldsymbol{\theta}^{(k)}\right)$.
- However, $f\left(\mathbf{y}_{\text{miss}},\boldsymbol{\tau}\mid\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\right)$ is very complicated, and we cannot even obtain $\bar{\mathbf{s}}\left(\boldsymbol{\theta}^{(k)}\right)$ in closed form. 😱
- We can take a stochastic approximation of the expectation but, still, drawing samples from $f\left(\mathbf{y}_{\text{miss}},\boldsymbol{\tau}\mid\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\right)$ is very complicated! 😱

Estimation of parameters for AR(1) Student's t

- We will use a Markov chain Monte Carlo (MCMC) process.

- In particular, we consider the Gibbs sampling method to generate the Markov chain: we divide the latent variables $(\mathbf{y}_{\text{miss}}, \boldsymbol{\tau})$ into two blocks $\boldsymbol{\tau}$ and $\mathbf{y}_{\text{miss}}$ and then generate a Markov chain of samples from the conditional distributions $f\left(\boldsymbol{\tau}|\mathbf{y}_{\text{miss}}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(k)}\right)$ and $f\left(\mathbf{y}_{\text{miss}}|\boldsymbol{\tau}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(k)}\right)$ alternatively:

  👉 *Drawing from* $f\left(\boldsymbol{\tau}|\mathbf{y}_{\text{miss}}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(k)}\right)$ *is trivial since the elements of* $\boldsymbol{\tau}$ *are iid following a univariate gamma distribution.*

  👉 *Drawing from* $f\left(\mathbf{y}_{\text{miss}}|\boldsymbol{\tau}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{(k)}\right)$ *is trivial since it's just a Gaussian distribution.*

Algorithm

Stochastic EM algorithm for AR(1) Student's t:

Initialize latent variables and set $k = 0$.

repeat

- **Simulation step**: generate the samples $\left(\boldsymbol{\tau}^{(k,l)}, \mathbf{y}_{\mathrm{m}}^{(k,l)}\right)$ $(l = 1, 2 \ldots, L)$ from $f\left(\mathbf{y}_{\mathrm{miss}}, \boldsymbol{\tau} | \mathbf{y}_{\mathrm{obs}}; \boldsymbol{\theta}^{(k)}\right)$ via Gibbs sampling (L parallel chains).
- **Approximation step**:

$$\hat{\mathbf{s}}^{(k)} = \hat{\mathbf{s}}^{(k-1)} + \gamma^{(k)} \left(\frac{1}{L} \sum_{l=1}^{L} \mathbf{s}\left(\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{miss}}^{(k,l)}, \boldsymbol{\tau}^{(k,l)}\right) - \hat{\mathbf{s}}^{(k-1)} \right).$$

- **Maximization step**: $\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\arg\max}\ \hat{Q}\left(\boldsymbol{\theta}, \hat{\mathbf{s}}^{(k)}\right)$.

$k \leftarrow k + 1$

until convergence

Maximization step

The maximization step $\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \hat{Q}\left(\boldsymbol{\theta}, \hat{\mathbf{s}}^{(k)}\right)$ can be obtained in **closed form**:

$$\varphi_0^{(k+1)} = \frac{\hat{s}_5^{(k)} - \varphi_1^{(k+1)}\hat{s}_7^{(k)}}{\hat{s}_3^{(k)}},$$

$$\varphi_1^{(k+1)} = \frac{\hat{s}_3^{(k)}\hat{s}_6^{(k)} - \hat{s}_5^{(k)}\hat{s}_7^{(k)}}{\hat{s}_3^{(k)}\hat{s}_4^{(k)} - \left(\hat{s}_7^{(k)}\right)^2},$$

$$\begin{aligned}\left(\sigma^{(k+1)}\right)^2 =& \frac{1}{T-1}\left(\hat{s}_2^{(k)} + \left(\varphi_0^{(k+1)}\right)^2\hat{s}_3^{(k)} + \left(\varphi_1^{(k+1)}\right)^2\hat{s}_4^{(k)}\right.\\ &\left. - 2\varphi_0^{(k+1)}\hat{s}_5^{(k)} - 2\varphi_1^{(k+1)}\hat{s}_6^{(k)} + 2\varphi_0^{(k+1)}\varphi_1^{(k+1)}\hat{s}_7^{(k)}\right),\end{aligned}$$

$$\nu^{(k+1)} = \underset{\nu>0}{\arg\max}\ (T-1)\left\{\frac{\nu}{2}\log\left(\frac{\nu}{2}\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right\} + \frac{\nu\hat{s}_1^{(k)}}{2}.$$

Convergence

The previous algorithm is very simple but does it converge? 🤔

Theorem:

The sequence $\left\{\boldsymbol{\theta}^{(k)}\right\}$ generated by the algorithm has the following asymptotic property: with probability 1, $\lim_{k\rightarrow+\infty} d\left(\boldsymbol{\theta}^{(k)}, \mathcal{L}\right) = 0$, where $d\left(\boldsymbol{\theta}^{(k)}, \mathcal{L}\right)$ denotes the distance from $\boldsymbol{\theta}^{(k)}$ to the set of stationary points of observed data log-likelihood $\mathcal{L} = \left\{\boldsymbol{\theta} \in \Theta, \frac{\partial l(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})}{\partial \boldsymbol{\theta}} = 0\right\}$.[a]

[a] J. Liu, S. Kumar, and D. P. Palomar, "Parameter estimation of heavy-tailed AR model with missing data via stochastic EM," *arXiv preprint*, 2018. [Online]. Available: https://arxiv.org/pdf/1809.07203.pdf.

Outline

Step 2: Imputation of missing values

- Given the conditional distribution $f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}})$, it is trivial to randomly generate the missing values (multiple realizations can be drawn for multiple imputation):

$$\mathbf{y}_{\text{miss}} \sim f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}}).$$

- However, in our case, we don't have the conditional distribution $f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}})$ in closed form:

$$f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}}) = \int f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) \mathrm{d}\boldsymbol{\theta}$$

- An *improper* way of imputing (which is acceptable in many cases with small percentage of missing values) is with $f(\mathbf{y}_{\text{miss}}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{\text{ML}})$, but even that expression is not available.
- We can instead draw from $f(\mathbf{y}_{\text{miss}}, \boldsymbol{\tau}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta}^{\text{ML}})$ and discard $\boldsymbol{\tau}$, but that expression is not available either. 😞

👉 *We can generate the samples from the joint based on Markov chains.*

Step 2: Imputation of missing values

- In particular, we consider the Gibbs sampling method to generate the Markov chain: we divide the latent variables $(\mathbf{y}_{\text{miss}}, \boldsymbol{\tau})$ into two blocks $\boldsymbol{\tau}$ and $\mathbf{y}_{\text{miss}}$ and then generate a Markov chain of samples from the conditional distributions $f(\boldsymbol{\tau}|\mathbf{y}_{\text{miss}}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$ and $f(\mathbf{y}_{\text{miss}}|\boldsymbol{\tau}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$ alternatively.

  👉 *Drawing from $f(\boldsymbol{\tau}|\mathbf{y}_{\text{miss}}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$ is trivial since the elements of τ are iid, so it is just a univariate gamma distribution for each element.*

  👉 *Drawing from $f(\mathbf{y}_{\text{miss}}|\boldsymbol{\tau}, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta})$ is just a Gaussian distribution.*

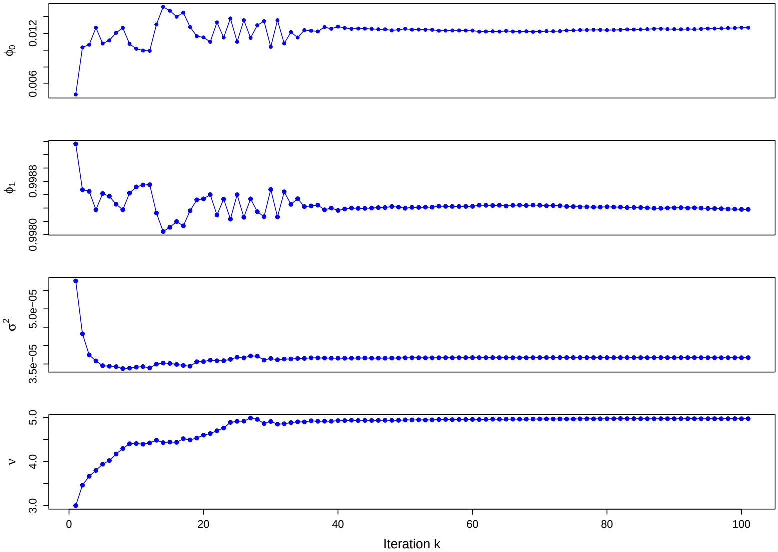
- If multiple imputation is needed, then the Markov chain has to be generated multiple times. But this is not the correct way to do multiple imputation! 😱

  👉 *The correct way is via a Bayesian characterization of $\boldsymbol{\theta}$ instead a point estimation like ML.*

Outline

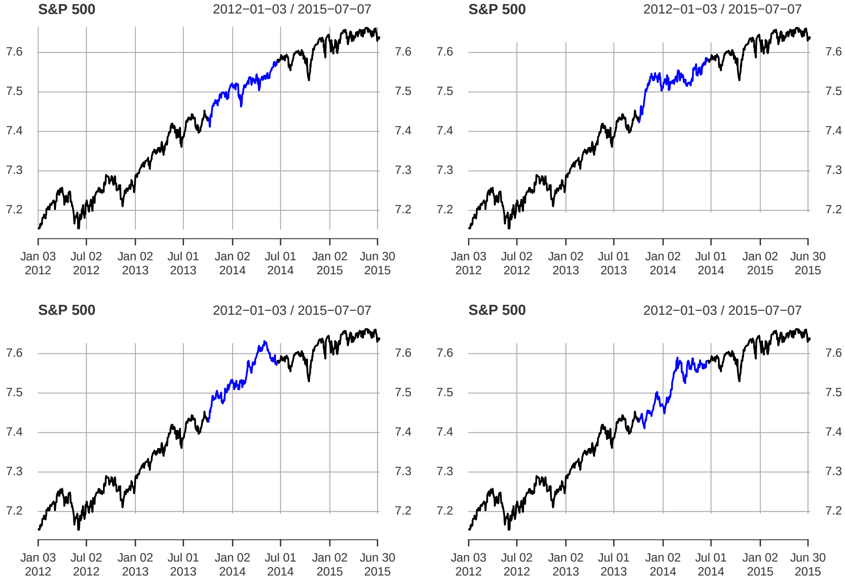
Numerical simulations

Estimation of parameters for AR(1) Student's t with real data (S&P 500):



Numerical simulations: imputed or real?

Numerical simulations: imputed or real?



Outline

Summary

- We have introduced the issue of missing values in observations.
- Imputation is the mechanism by which one fills in those missing values.
- Many methods are ad-hoc with no good statistical results.
- Other methods are based on some properly defined formulation based on some structural properties of the data matrix.
- Time series contain special temporal structure that can be employed for imputation.
- Sound statistical method:
  1. estimate the statistics of the underlying distribution function and construct the conditional distribution
  2. impute based on the conditional distribution either a single time or multiple times (multiple imputation)

Thanks

For more information visit:

https://www.danielppalomar.com