Analysis and Synthesis of Pseudo-Periodic Job Arrivals in Grids: A Matching Pursuit Approach

Hui Li∗
Leiden Institute of Advanced
Computer Science, Leiden University
PO Box 9512, 2333 CA, Leiden,
The Netherlands

Richard Heusdens
Delft University of Technology
Dept. of Mediamatics
Mekelweg 4, 2628 CD, Delft,
The Netherlands

Michael Muskulus
Mathematical Institute
Leiden University
PO Box 9512, 2333 CA, Leiden,
The Netherlands

Lex Wolters
Leiden Institute of Advanced
Computer Science, Leiden University
PO Box 9512, 2333 CA, Leiden,
The Netherlands

Abstract

Pseudo-periodicity is one of the basic job arrival patterns on data-intensive clusters and Grids. In this paper, a signal decomposition methodology called matching pursuit is applied for analysis and synthesis of pseudo-periodic job arrival processes. The matching pursuit decomposition is well localized both in time and frequency, and it is naturally suited for analyzing non-stationary as well as stationary signals. The stationarity of the processes can be quantitatively measured by permutation entropy, with which the relationship between stationarity and modeling complexity is excellently explained. Quantitative methods based on the power spectrum are also provided to measure the degree of periodicity present in the data. Matching pursuit is further shown to be able to extract patterns from signals, which is an attractive feature from a modeling perspective. Real world workload data from production clusters and Grids are used to empirically evaluate the proposed measures and methodologies.

1 Introduction

Production Grids nowadays consist of many clustered resources and scheduling in such an environment should be carried out at multiple levels. There are local resource management systems on clusters, Grid-level brokers, and VO-based schedulers. Consequently, performance evaluation of different scheduling strategies requires representative workload models [2]. It is firstly reported in [9] that strong pseudo-periodic components are present in job interarrival time processes for certain VOs at the Grid level, which makes it difficult to model such processes with stochastic models. Pseudo-periodicity is further categorized in [8] as one of the basic job arrival patterns on data-intensive clusters and Grids. In this paper we focus on the analysis and synthesis of pseudo-periodic job arrival processes.

The contribution of this work is three-fold. Firstly, job arrivals are studied as point processes and it is shown that measures based on interarrivals are of limited usefulness and count based measures should be trusted instead in revealing the periodic patterns. Secondly a signal decomposition methodology called matching pursuit is adapted and applied successfully in analyzing and modeling job arrivals. Moreover, matching pursuit is shown to be able to extract patterns from signals and makes it possible to model patterns separately. Finally, methods are provided to quantitatively measure the degree of periodicity in the data and permutation entropy is proposed for studying the stationarity of processes. The relationship between stationarity and modeling complexity is explained and a deeper understanding is obtained for the proposed matching pursuit approach.

The rest of this paper is organized as follows. Section 2 discusses point processes and their representations. Section 3 defines the periodicity measures and explains the origin of pseudo-periodic job arrival patterns to our best knowledge. Section 4 introduces basic notions such as atoms and dictionaries and outlines a standard iterative matching pursuit algorithm. Section 5 presents the ex-
Job Arrivals as Point Processes

Job arrivals can be described as a point process, which is defined as a mathematical construct that represents individual events as random points at times \( \{ t_n \} \). There are different representations of a point process. The counting process \( \{ N(t) \} \) is a continuous-time, non-negative stochastic process where \( N(t) = \max\{ n : t_n \leq t \} \) is the number of arrivals in the interval \((0,t]\). The interarrival time process \( \{ I_n \} \) is a real-valued random sequence with \( I_n = t_n - t_{n-1} \). The counting process and the interarrival process are stochastically equivalent and either of them completely describe the point process.

The sequence of counts, or the count process, is formed by dividing the time axis into equally spaced contiguous intervals of \( T \) to produce a sequence of counts \( \{ C_k(T) \} \), where \( C_k(T) = N((k+1)T) - N(kT) \) denotes the number of events in the \( k \)th interval. This sequence forms a discrete-time random process of non-negative integers and it is another useful representation of a point process. The rate process \( R_k(T) \) is a normalized version of the sequence of counts, where \( R_k(T) = C_k(T)/T \).

In general, forming the sequence of counts loses information because interarrivals between events within the interval \( T \) are not preserved. Nevertheless, this representation is very important because it preserves the correspondence between the discrete time axis of the count process \( \{ C_k(T) \} \) and the absolute “real” time axis of the underlying point process. We can readily associate correlation in the process \( \{ C_k(T) \} \) with correlation in the point process. The interarrival time process, on the other hand, eliminates the direct correspondence between the absolute time and the index number therefore it only allows rough comparisons with correlations in the point process [10]. As is shown in [8], measures based on interarrivals are not able to reliably reveal the correlations of the underlying processes. The study of periodicity requires second-order statistics such as the autocorrelation function and power spectrum, so count/rate-based measures are the main focus in our empirical studies.

3 Pseudo-Periodicity

From Fourier analysis it is well known that periodicity shows up as peaks in the frequency domain. Real world data, however, seldom exhibits perfectly periodic behavior. In most situations pseudo-periodic signals are observed instead, potentially arising from various sources of noises and the time-varying nature of the generation scheme. From this perspective it is necessary to use quantitative methods to measure the degree of periodicity in the data. We provide two simple measures for this purpose. The first periodicity measure \( P_f \) is defined as the normalized difference of the sum of the power spectrum values at the highest amplitude frequency and its multiples, and the sum of the power spectrum values at the halfway-between frequencies [13]. The second measure is defined as the saturated entropy of the signal power spectrum [4]. Both measures have values between 0 and 1. Higher \( P_f \) and lower entropy correspond to stronger periodicity in the signal. As is shown in Section 5 these measures give a rough indication of the strength of periodicity in signals.

Figure 1 and 2 show the first and second order statistics for the job interarrival time process and one corresponding count process of a typical pseudo-periodic process. In Figure 1 a strong deterministic component is observed in the histogram plot of job interarrivals and periodicity is detected as the multimodals in the low frequency domain. As to the count process (scale\(^1\)=6) shown in Figure 2, periodicity is clearly observed in the time series of the signal and by the strong harmonic peaks in the power spectrum. It has to be noted that there are cases in which no obvious periodic components are found in the interarrival times but the count process exhibits periodic behavior. In most cases count based measures should be trusted to reliably reveal the correlation structures of the underlying point process.

\(^1\)A dyadic scale is used such that scale \( j \) means interval \( T = 2^j \) seconds in the count based measures.
Our focus in this paper is on data-intensive Grid environments whose workloads consist of massive single-CPU jobs. We try to understand the origin of pseudo-periodic job arrivals in such environments. The example used in this section for the pseudo-periodic pattern is called lhcb, which is a large-scale high energy physics experiment in the LCG Grid. If we take into account that close to 90\% of lhcb jobs (around 60,000) are from a single “user” during the eleven consecutive days under study, we can assume that scripts are used to submit those jobs, which are largely deterministic in nature. It can also be assumed that automated tasks need to be implemented to process such a huge amount of scientific data. Periodic behavior can also originate from testing and monitoring jobs in the Grid such as those from dteam. The function of dteam (“deployment team”) is to assure a continuously operating Grid. Mostly testing and monitoring jobs are initiated automatically by software in a pseudo-periodic fashion. The periodic pattern is also observed for VO’s at the cluster level [8]. We consider it as one basic pattern that originates from automated submission schemes, which is present in large-scale data-intensive environments. Being able to model such patterns is an indispensable part of comprehensive workload modeling in Grids, which has important implications in experimental performance studies. Our approach for modeling pseudo-periodicity in Grid job arrivals is inspired and adapted from methodology in sinusoidal modeling of audio signals, and it is elaborated in the following section.

4 Matching Pursuit

Sinusoidal modeling has been widely used in modeling pseudo-periodic signals, especially in audio signal processing [3, 12]. The sinusoidal parameters can be estimated by methods such as spectral peak picking and analysis-by-synthesis. We focus on a particular analysis-by-synthesis method called matching pursuit. The matching pursuit algorithm is introduced in [11] for signal decomposition and several variations have been proposed [3, 5, 6, 7]. It is a greedy, iterative algorithm which searches a set of candidate functions for the element that best matches the signal and subtracts this function to form a residual signal to be approximated in the next iteration. The following subsections include a paraphrase of [5] to introduce the basic notions and a standard matching pursuit algorithm.

4.1 Atoms and Dictionaries

Decompositions of signals over families of functions that are well localized both in time and frequency have important applications in signal processing and beyond. Windowed Fourier transforms and wavelet transforms are well-studied examples of such decompositions. The basic functions or waveforms are called time-frequency atoms. Gabor atoms are a general family of waveforms that are widely used for decomposition and they have the form

\[ g_{s,u,\xi}(t) := \frac{1}{\sqrt{s}} w\left(\frac{t-u}{s}\right) e^{i2\pi \xi(t-u)}, \]

where \( s, \xi, u \) represent scale, frequency, and translation, respectively. Gabor atoms are obtained by dilating, translating, and modulating a mother window \( w(t) \), which is real-valued, positive, and satisfies \( \int |w(t)|^2 dt = 1 \). The energy of a particular Gabor atom is centered at time \( u \) and is proportional to \( s \). Fourier transforming \( g_{s,u,\xi}(t) \) results in \( g_{s,u,\xi}(f) \), whose energy is concentrated around \( \xi \) with a size proportional to \( 1/s \).

Given the strong harmonic components in signals, it is natural to define harmonic atoms as

\[ h(t) := \sum_{k=1}^{K} c_k g_{s_k,u_k,\xi_k}(t), \int |h(t)|^2 dt = 1, \]

where \( \xi_k \approx k \xi_0, c_k \) are complex coefficients, \( 1 \leq k \leq K \). Compared with a Gabor atom, the Fourier transform of a harmonic atom has \( K \) peaks located around frequencies \( \xi_k \) with a common size proportional to \( 1/s \).

Real-valued Gabor atoms are of the form

\[ g_{s,u,\xi,\phi}(t) := c_{s,\xi,\phi} w\left(\frac{t-u}{s}\right) \cos(2\pi \xi(t-u) + \phi), \]

where \( \phi \) represents the phase of the cosine function. Gabor atoms are special cases of harmonic atoms.

A dictionary represents a redundant collection of basic waveforms. For instance, a Gabor dictionary is a set \( \mathcal{D}_g = \{g_{s,u,\xi} \in \mathbb{R} = \mathbb{R} \times \mathbb{R}^2 \} \). A harmonic dictionary \( \mathcal{D}_h \) is an extension of the Gabor dictionary \( \mathcal{D}_g \) [5]. Signals are decomposed into a linear expansion of waveforms that belongs to a redundant dictionary via matching pursuit.

4.2 The Standard Matching Pursuit

A matching pursuit is a greedy algorithm that chooses at each iteration a waveform that is best adapted to approximate a part of the signal. Given a complete dictionary \( \mathcal{D} \) and a number \( M > 0 \), it decomposes a signal \( s(t) \) into a residual part \( R_M(t) \) and a linear combination of \( M \) atoms chosen from \( \mathcal{D} \)

\[ s(t) = \sum_{m=1}^{M} \lambda_m g_m(t) + R_M(t), \quad ||g_m|| = 1, \]

with the energy conservation property

\[ ||s||^2 = \sum_{m=1}^{M} \lambda_m^2 ||g_m||^2 + ||R_M||^2. \]

Given the strong convergence \( \lim_{M \to \infty} ||R_M|| = 0, \) it is proven that an arbitrarily good approximation to a signal \( s(t) \) can be obtained.
A standard iterative matching pursuit can be described as Algorithm 1. A fast matching pursuit algorithm with harmonic dictionaries is discussed in detail in [5]. We make use of the core implementation in MPTK [7] and adapt the code to analyze and synthesize Grid job arrival processes.

Algorithm 1 Standard Matching Pursuit

1: Initialization: $M = 0$, $R_0(t) = s(t)$;
2: while the number of extracted atoms $M$ is less than the desired number or the signal-to-noise ratio (SNR) has not yet reached the predefined level do
3:   $\forall g \in D$, compute $||R_M, g||$;
4:   Select the best matched atom from the dictionary as $g_{M+1}$: $||R_M, g_{M+1}|| \geq \sup_{g \in D} ||R_M, g||$;
5:   Update the residual $R_{M+1}(t) := R_M(t) - \langle R_M, g_{M+1} \rangle g_{M+1}(t)$.
6: end while
7: $\hat{s}(t) = \sum_{m=0}^{M-1} \langle R_M, g_{m+1} \rangle g_{m+1}(t)$, residual $R_M(t)$.

5 Empirical Evaluation

In this section we present experimental studies on real world workload data which show pseudo-periodic job arrival patterns. Table 1 presents a summary of workload traces under study. LCG1 and LCG2 are two traces from the LHC Computing Grid. The LCG production Grid currently has approximately 180 active sites with around 30,000 CPUs and 3 petabytes storage, which is primarily used for high energy physics data processing. There are also biomedical applications running on the Grid. Almost all the jobs are massively parallel tasks, requiring one CPU to process a certain amount of data. The workloads are obtained via the LCG Real Time Monitor\footnote{The Real Time Monitor is developed by Imperial College London and it monitors jobs from most of the major Resource Brokers in LCG, therefore the data it collects is representative at the Grid level. http://gridportal.hep.ph.ic.ac.uk/rtm/} for two periods: LCG1 consists of jobs of eleven consecutive days from November 20th to 30th in the same year. LCG2 is from December 19th to 30th in the same year. As explained in Section 3, lhcb and dteam are the two main VO's that exhibit pseudo-periodic behavior. atlas on LCG2, another large HEP experiment, contains certain high frequency component at small scales ($\leq 6$) and it is used as an example to show pattern extraction with matching pursuit. At the cluster level we use traces from two data-intensive clusters, namely, NIK05 and LPC05. They are located at the HEP institutes in the Netherlands, and France, respectively, and both participate in LCG. It should be noted that these clusters are involved in multiple different collaborations and have their own local user activities. com1 is a company partner with NIKHEF which runs medical-related data-intensive jobs. biomed is the VO with biomedical applications and it contributes to $\sim 65\%$ of all LPC05 jobs. Job arrivals in biomed are not periodic at all but long range dependent (LRD), which is included here to demonstrate how a matching pursuit decomposition can be used to approximate arbitrary signals.

Firstly the periodicity measures are applied to the data under study and the results are shown in Table 2. The total spectrum entropy (TSE) calculates the entropy for the whole power spectrum while the saturated spectrum entropy (SSE\footnote{Due to the pronounced zero-frequency peak of the spectrum, the entropy of the spectrum, when cut off at a lower frequency $f_0$, usually rises when $f_0$ is raised, and saturates quickly to the value we call SSE. For long-range dependent processes this is not the case (due to their particular infrared behaviour), such that SSE does not exist for them.}) excludes the first one or two “big” power spectrum values, which represent the total energy of the signal. SSE values should be used to examine the strength of periodicity and its results are consistent with those of $P_f$: lower SSE values correspond to higher $P_f$ values, which indicate stronger periodic behavior. We can see that all processes (except biomed, LPC05) show quite strong periodicity. Among all dteam exhibits the strongest periodic behavior with $P_f$ reaching 0.95. These simple measures give us a rough idea about the degree of periodicity in the signal and more sophisticated methods should be investigated for a more strict treatment.

In the following subsections we investigate extensively the applicability of matching pursuit in the analysis and synthesis of pseudo-periodic job arrivals. Firstly the stationarity of the signal is studied using the short-time Fourier transform and the permutation entropy. It is shown that for pseudo-periodic signals the complexity of a matching pursuit decomposition is directly related to the stationarity of the process (as measured). Secondly, we study in detail the properties of the matching pursuit approach: the number of atoms needed, signal-
to-noise ratio (SNR)\(^4\) achieved, and heuristics for a good stop criterion from a modeling perspective. Finally, we demonstrate that matching pursuit is a powerful tool in extracting patterns from signals, which makes it possible to model these patterns individually.

5.1 Stationarity

The short-time Fourier transform (STFT) is a simple way to show time and frequency information simultaneously by Fourier transforming signals by small windows over time. Figure 3 shows the STFTs of three signals via spectrogram. The simulated “three-cosine” signal, consisting of a superposition of three different cosines, is obviously stationary with the same harmonic components over time. The job arrival count process of \(lhcb\), \(LCG1\), however, contains different frequency contents in different periods. Nevertheless, its energy concentrates mainly on the strong harmonic components which remain roughly the same over time. It indicates that \(lhcb\) stays “closer” to a stationary signal. \(dteam\), on the other hand, exhibits a much richer frequency content including the harmonics (especially in the first and the last quarter of the time axis). It shows that \(dteam\) is not as stationary as its \(lhcb\) counterpart.

Stationarity can be better analyzed and illustrated by a novel method called permutation entropy (PE). Here we only briefly introduce it for understanding our results and refer the readers to [1] for a thorough treatment. Permutation entropy is a complexity measure for times series analysis and it can be used to detect dynamic changes in signals. The degree of non-stationarity of a signal is reflected in a higher variability of its PE. Results for the data under study are shown in Figure 4. Clearly the “three-cosine” example is the most stationary by appearing as a straight line in the PE plot. \(lhcb\)'s are relatively more stationary with a smooth and slow-varying PE curve. \(dteam\)'s, on the other hand, are non-stationary with many jumps and abrupt changes. The rough idea about stationarity obtained via STFT is nicely quantified and verified by permutation entropy. The entropy values themselves can also be beautifully linked with the periodicity measures we studied above. \(lhcb\)'s are more stationary but less periodic (more stochastic components), while \(dteam\)'s are less stationary but more periodic (less stochastic components). The other two processes at the cluster level, namely \(com1\), \(NIK05\) and \(all, NIK05\), exhibit even more non-stationary behavior.

We argue that the the stationarity of a pseudo-periodic signal is directly related to the complexity of the matching pursuit decomposition, namely, how many atoms are used to reach a certain signal-to-noise ratio (SNR). Naturally a more stationary signal requires less atoms for a good approximation, which results in a simpler model. This relationship is empirically shown in the process of decomposing signals into sinusoidal components and residuals.

5.2 Signals and Residuals

Figure 5 shows the number of atoms used in the matching pursuit decomposition versus the corresponding SNR achieved. For the stationary “three-cosines”

\(^4\)SNR is defined as 10 times the decadic logarithm of the power ratio: \(SNR(dB) = 10\log_{10}(\frac{P_{signal}}{P_{noise}})\).
Table 2. The periodicity measures and the signal-to-noise ratio achieved by different number of atoms.

<table>
<thead>
<tr>
<th>Trace</th>
<th>TSE</th>
<th>SSE</th>
<th>P_t</th>
<th>N_a = 20</th>
<th>N_a = 50</th>
<th>N_a = 100</th>
<th>N_a = 500</th>
<th>N_a = 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lhcb, LCG1 (scale=6)</td>
<td>0.40</td>
<td>0.74</td>
<td>0.84</td>
<td>10.99</td>
<td>12.64</td>
<td>14.73</td>
<td>25.52</td>
<td>36.30</td>
</tr>
<tr>
<td>lhcb, LCG2 (scale=6)</td>
<td>0.18</td>
<td>0.72</td>
<td>0.78</td>
<td>13.95</td>
<td>15.81</td>
<td>17.74</td>
<td>27.60</td>
<td>36.69</td>
</tr>
<tr>
<td>dteam, LCG1 (scale=6)</td>
<td>0.69</td>
<td>0.71</td>
<td>0.94</td>
<td>5.82</td>
<td>8.51</td>
<td>11.30</td>
<td>24.45</td>
<td>36.56</td>
</tr>
<tr>
<td>dteam, LCG2 (scale=6)</td>
<td>0.68</td>
<td>0.70</td>
<td>0.95</td>
<td>6.09</td>
<td>9.11</td>
<td>12.11</td>
<td>25.01</td>
<td>36.14</td>
</tr>
<tr>
<td>com1, NIK05 (scale=8)</td>
<td>0.79</td>
<td>0.80</td>
<td>0.89</td>
<td>2.52</td>
<td>3.21</td>
<td>4.03</td>
<td>9.03</td>
<td>14.72</td>
</tr>
<tr>
<td>all, NIK05 (scale=8)</td>
<td>0.88</td>
<td>0.91</td>
<td>0.79</td>
<td>1.54</td>
<td>2.15</td>
<td>2.88</td>
<td>5.79</td>
<td>8.25</td>
</tr>
<tr>
<td>biomed, LPC05 (scale=8)</td>
<td>0.63</td>
<td>-</td>
<td>0</td>
<td>3.44</td>
<td>4.22</td>
<td>4.81</td>
<td>7.05</td>
<td>8.92</td>
</tr>
</tbody>
</table>

Figure 5. 1/SNR against the number of atoms.

Figure 7. SNR against the STFT window size.

data and the close-to-stationary lhcb processes, satisfactory SNRs are reached after very limited iterations (i.e. two) and then increase only slowly with more atoms. For all three processes one constant atom and one harmonic atom reproduce the majority of energy of the original signal. This is in accordance with the fact that stationarity simplifies the fitted sinusoidal models. To closely examine how the synthesized signal resembles the original one with respect to the number of atoms, we use lhcb, LCG1 as an illustrative example. As is shown in Figure 6, the synthetic signal, the residual, and the residual spectrum are plotted for an increasing number of atoms from left to right. For atoms = 2, we can see that a basic “tune” is set by the synthesized signal. However, the residual still contains a significant part of the signal, especially in the higher frequency domain. In other words, the synthesis at this level lacks the “dynamics” present in the underlying signal. As the atoms increase to 6 and to 16, we can observe that more and more components in the original signal represented by atoms are added to the synthetic one, which makes the matching better. With 500 atoms the synthesis closely resembles the original one. The residual is white-noise like and contains very little energy. As is shown in Table 2 nearly perfect reconstruction is achieved for lhcb, LCG1 with around 1,000 atoms (SNR=36.30).

For dteam SNRs of the matching pursuit decomposition increase along with the number of atoms used, meaning that more components are needed for a better match. The achieved SNRs up to 100 atoms (see Table 2) are also lower than their lhcb counterparts because of their non-stationarity. Nevertheless, for number of atoms larger than 500 the matching pursuit performs similarly for both lhcb and dteam. At the cluster level for VO com1 and the aggregated whole trace NIKo5, more atoms are needed for reaching a certain SNR level and the absolute SNR values are considerably smaller than the Grid level VOIs. This is partially explained by the fact that the energy of the harmonic contents are relatively lower compared to that of lhcb or dteam. It may also be because that the arrival processes are less stationary at the cluster level, especially for the aggregated whole data trace. By increasing the number of atoms better approximations can be obtained but there is a tradeoff of increased complexity. In theory the matching pursuit can approximate arbitrarily close to a certain signal given that the number of atoms grows unlimitedly. This type of decomposition can be used to analyze and synthesize all types of signals besides periodic ones. For instance, results of matching are shown for biomed, LPC05 in Table 2, whose job arrivals form a long range dependent (LRD) process. It can be reproduced well by the matching pursuit with a sufficient number of atoms.

It is also interesting to investigate to what extent the matching of a non-stationary signal can be improved with a reduced window size of the basic function. Figure 7 shows the SNR versus the smallest window size used in the dictionary for different number of atoms. Generally speaking a redundant dictionary including smaller window sizes means that there are more “short” building blocks available for addition, which would result in a potential better fitting. For the close-to-stationary lhcb
A crucial issue in the matching pursuit practices is how to choose a good stop criterion. One of them is the signal-to-noise ratio (SNR). The iterative matching pursuit would proceed until a predefined SNR threshold is reached. Other measures have to be considered as well for an educated decision. From a modeling for performance evaluation perspective, a simple parametric model is preferable with controllable parameters. For a close-to-stationary case like \textit{lhcb} (see Figure 6), using 16 atoms is a good choice because the dynamics of the process is captured while the parameter space is still relatively small. There is a tradeoff between the number of atoms and the SNR in the decomposition process. The line between the signal and the residual is also blurred because parts of the signal are embedded in the residual if only a limited number of atoms are used for approximation. For this reason we do not study models for the residual (noise), as it is largely data dependent. If a large number of atoms can be used, however, it becomes a trivial task to model the residual because it resembles the uncorrelated or weakly-correlated Gaussian white noise with well-controlled energy. For more non-stationary situations it is up to the practitioner to decide when to stop. For the purpose of simulation studies of scheduling strategies, further research is needed to reveal the impact of both periodic and non-stationary parts in the arrival processes.

### 5.3 Pattern Extraction

Another very attractive feature of the matching pursuit is that it can be used to extract patterns from signals. Lots of real world signals contain multiple patterns and sometimes it is desirable to separate and model them individually. Figure 8 shows the job arrival count process of \textit{atlas}, \textit{LCG2} at scale 6. We can see that the process itself is long range dependent (LRD) with certain high frequency component. It is identified by the small oscillations along with the ACF decay and the peak in the power spectrum. This pattern is well extracted from the original signal by the matching pursuit and it is shown in the ACF plot and spectrum for the remaining part. It is possible to approximate the remaining signal by long

---

**Figure 6.** Plots of the reconstruction and the residual signal with an increasing number of atoms.
range dependent models. The extracted signal and its spectrum are illustrated, too. It is observed that the high frequency pattern occurs only at specific periods in time, a non-stationary situation for which the matching pursuit works naturally well.

6 Conclusions and Future Work

In this paper we studied the pseudo-periodic job arrivals on data-intensive clusters and Grids. Matching pursuit is successfully applied for analysis and synthesis of such processes. We find that the stationarity of a signal is directly connected to the complexity of modeling. Non-stationary signals generally require more atoms to reproduce their dynamics. Grid level processes such as lhcb and dteam are more stationary than the arrival processes at the cluster level, and the energy of their harmonic content is much stronger. This is evidenced by the smaller number of atoms used for synthesis and their higher SNRs achieved. For a close-to-stationary pseudo-periodic signal, a simple sinusoidal or harmonic model can be fitted with a manageable number of parameters by matching pursuit. The matching pursuit is also shown as a powerful tool to extract patterns from signals, which is useful to decompose complex signals and model the parts individually.

Our ongoing research includes the investigation of the impact of both periodic components and non-stationary parts on the evaluation of Grid-level scheduling strategies. It can help to determine a better stop criterion and potentially reduce the complexity of modeling. We are also preparing to release the adapted matching pursuit software as part of the Grid workload modeling tools5.

Acknowledgments

The LCG Grid traces are provided by the HEP e-Science group at Imperial College London. The NIK05 trace is provided by NIKHEF and LPC05 is obtained from Parallel Workload Archive. We want to thank all who graciously provide us with the data. The research of M. Muskulus has been funded by NWO grant no. 635.100.006.

References