Seismological Research Letters

Towards the processing of large data volumes with phase cross-correlation

--Manuscript Draft--

Manuscript Number:	SRL-D-19-00022R2
Full Title:	Towards the processing of large data volumes with phase cross-correlation
Article Type:	Electronic Seismologist
Corresponding Author:	Sergi Ventosa Institut de Ciències de la Terra Jaume Almera (ICTJA-CSIC) Barcelona, SPAIN
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Institut de Ciències de la Terra Jaume Almera (ICTJA-CSIC)
Corresponding Author's Secondary Institution:	
First Author:	Sergi Ventosa
First Author Secondary Information:	
Order of Authors:	Sergi Ventosa
	Martin Schimmel
	Eleonore Stutzmann
Order of Authors Secondary Information:	
Manuscript Region of Origin:	SPAIN
Suggested Reviewers:	Brigitte Knapmeyer-Endrun, Dr. Head of Seismological Observatory, University of Cologne brigitte.knapmeyer-endrun@uni-koeln.de
	Benoît Tauzin, Dr. Associate Professor, Université de Lyon benoit.tauzin@univ-lyon1.fr
	Clifford H. Thurber, Prof. University of Wisconsin–Madison cliff@geology.wisc.edu
	Helle Pedersen, Dr. Physicien CNAP, Université Grenoble Alpes, ISTerre helle.pedersen@univ-grenoble-alpes.fr
	Lise Retailleau, Dr. PostDoc, Stanford retaille@stanford.edu
	Philippe Jousset, Dr. Senior Researcher, Helmholtz Centre GFZ philippe.jousset@gfz-potsdam.de
	Paulssen Hanneke, Dr. Assistant Professor, Faculty of Geosciences, Utrecht University h.paulssen@uu.nl

Dear Editor,

The phase cross-correlation (PCC) method contributes to increase convergence of interstation correlations to a robust signal extraction in seismic ambient noise imaging and monitoring applications. We introduce solutions to severely reduce the computational cost of PCC. Further, we extract Rayleigh and body wave signals from many years of data from seismic stations distributed worldwide.

Thank you for considering this manuscript for publication,
Sergi Ventosa, Martin Schimmel and Eleonore Stutzmann

Response to reviews of "Toward the processing of large data volumes with phase cross-correlation" by Ventosa, Schimmel and Stutzmann

Comments from the Editors and Reviewers:

1. Please double check to make sure that the figures and references are properly cited (In the current clean version, they are filled with ??).

Solved.

2. Figure 3: please label the top and bottom panels as (a) and (b), and refer them in the figure caption.

We have labeled subfigures a and b, and referred them in the figure caption and the text.

3. Line 291: change "We thank to the editor and ..." to "We thank the editors and .."?

Done.

Towards the processing of large data volumes with phase cross-correlation

Sergi Ventosa*, Martin Schimmel*, Eleonore Stutzmann†

Abstract

Interstation correlation is the basic operation in seismic noise and coda wave interferometry for signal extraction in imaging and monitoring applications. Conventional cross-correlations evaluate the similarity between two signals along lag time and they are efficiently computed by the fast Fourier transform (FFT), valuable to manage the large data volumes that ambient noise applications demand. The phase cross-correlation (PCC) method contributes to increase convergence, a key issue in seismic ambient noise imaging and monitoring; however, it is much more computationally demanding. PCC evaluates similarity by subtracting the modulus of the sum and difference of the instantaneous phase of two signals. We introduce solutions to dramatically reduce the high computational cost of PCC. We show that PCC can be rewritten as a complex cross-correlation and computed by the FFT when the moduli are raised to the power of 2, and we demonstrate PCC can improve waveform coherence and increase convergence compared to the default processing flow of 1-bit amplitude normalization and standard cross-correlation. Moreover, we develop a graphics processing unit implementation to accelerate computations when using powers other than 2 and particularly

*Institute of Earth Sciences Jaume Almera - CSIC, E-08028 Barcelona, Spain.

†Institut de Physique du Globe de Paris, Sorbonne Paris Cité, Univ Paris Diderot, UMR 7154 CNRS, F-75005 Paris, France.

19 when using the power of 1. Finally, we extract Rayleigh and body wave signals from many
20 years of data from seismic stations distributed worldwide using PCC without a significant
21 increase in computational cost compared to conventional cross-correlation.

22 Introduction

23 Cross-correlation is a ubiquitous method in seismic data processing important for picking,
24 empirical Green’s function extraction, ambient noise monitoring, waveform comparison, and
25 signal, event and pattern detection or identification among others. Phase cross-correlation
26 (PCC) was conceived to better detect weak P-to-S conversions at the upper mantle disconti-
27 nities (Schimmel, 1999). In this work, PCC is introduced to estimate travel-time differences
28 by searching for the lag time that maximize an amplitude-unbiased misfit function between
29 the pilot and converted waveforms, their phase coherency. Phase coherency is useful for
30 detecting weak signals in the coda of stronger waves because it effectively considers a larger
31 portion of the waveforms than conventional correlation-based methods that consider mainly
32 the most energetic components. Later on, PCC was found useful to improve waveform coher-
33 ence and convergence of interstation correlations thanks to its amplitude-unbiased property
34 (Schimmel *et al.*, 2011). Essentially, PCC can be useful for applications demanding more
35 information on the waveform.

36 Interstation correlations of seismic ambient noise and scattered coda waves complement
37 direct arrivals from earthquakes in surface-wave tomography, monitoring, and Earth’s deep
38 interior studies in regions otherwise poorly illuminated. Convergence to the empirical Green’s
39 function or a robust seismic signal extraction increases compared to the direct correlation of
40 raw seismic records through customized processing flows that improve the balance of noise
41 sources and reduce the influence of anomalous signals. These processing flows compute many
42 correlations on relatively short data sequences. The pre-processing of the raw data may in-
43 clude instrument response correction, anomalous signal rejection, spectral whitening, and

time-domain normalization (Bensen *et al.*, 2007), and once the correlations are computed, weighted stacking and denoising (Baig *et al.*, 2009; Schimmel *et al.*, 2011; Cheng *et al.*, 2015; Moreau *et al.*, 2017). Finally, subsequent measurements of group and phase velocity dispersion are done in surface-wave tomography applications (e.g., Haned *et al.*, 2016; Schippkus *et al.*, 2018; Obermann *et al.*, 2016), of velocity and structure variations in monitoring (e.g., Hadziioannou *et al.*, 2011; D’Hour *et al.*, 2016; Takano *et al.*, 2017; Sánchez-Pastor *et al.*, 2018), or of travel times of body waves from ambient noise or earthquake coda in Earth’s deep interior studies (e.g., Nishida, 2013; Boué *et al.*, 2014; Phạm *et al.*, 2018; Tkalčić and Phạm, 2018).

Spectral whitening or band-pass filtering to the frequency band of interest followed by 1-bit amplitude normalization and geometrical-normalized cross-correlation (GNCC) is the processing flow most often used, where GNCC is the conventional cross-correlation divided by the standard deviation of the signals. The 1-bit cross-correlation method applies correlation on the sign of the input sequences (e.g., Bensen *et al.*, 2007; Cupillard *et al.*, 2011). This instantaneous amplitude normalization reduces the influence of poorly-distributed high-energy features such as earthquakes but it distorts waveforms and slows down convergence (Schimmel *et al.*, 2018). The PCC method measures similarity between two analytic signals of unitary amplitude, i.e., two instantaneous phases. Schimmel *et al.* (2011) show that PCC is a less aggressive alternative to 1-bit cross-correlation that may help to improve waveform coherence and increase convergence. This has been confirmed, for example, in D’Hour *et al.* (2016) for monitoring of hydromechanical changes, in Taylor *et al.* (2016) and Becker and Knapmeyer-Endrun (2018) for retrieving P wave Moho reflections from noise autocorrelations, in Haned *et al.* (2016) for extracting group-velocity observations for global tomography using seismic hum, in Schimmel *et al.* (2018) for extracting Rayleigh wave group velocities using phase autocorrelation, and in Romero and Schimmel (2018) for mapping the basement of the Ebro Basin in Spain using shallow subsurface reflections.

The computational cost of the signal processing methods conventionally used in the pre-

71 processing of the raw time series is small compared to the correlation, weighted stacking,
 72 and denoising operations subsequently applied. This is mainly, because pre-processing is
 73 done per station and correlation and denoising operations are done per station pair. Ob-
 74 sPy (Krischer *et al.*, 2015) and MSNoise (Lecocq *et al.*, 2014) are open-source packages for
 75 Python that are excellent for proof-of-concept studies and often sufficient for pre-processing.
 76 Compiled programming language solutions using the central processing unit (CPU) or the
 77 graphics processing unit (GPU) are however necessary in order to apply the most demand-
 78 ing operations to large data volumes in a reasonable amount of time. Several tools have
 79 been introduced recently for improving the computational efficiency of key operations along
 80 interstation correlation processing flows. Zeng and Thurber (2016) present a GPU imple-
 81 mentation of the time-frequency phase-weighted stack (tf-PWS; Schimmel *et al.*, 2011) which
 82 is up to 20 times faster than equivalent CPU implementations using the FFT. Ventosa *et al.*
 83 (2017) reduce computational cost of tf-PWS by introducing the time-scale phase-weighted
 84 stack (ts-PWS) that uses the wavelet transform for the time-frequency representation, and
 85 propose unbiased phase coherence and two-stage ts-PWS to improve convergence speed and
 86 quality of the extracted signals. The unbiased phase coherence estimator contributes to
 87 increase non-coherent noise attenuation because it equals to 0 if the signals are totally inco-
 88 herent. Whereas the two-stage ts-PWS reduces signal attenuation by stacking linearly the
 89 sequences in a few groups to improve signal-to-noise ratio before applying the actual ts-PWS
 90 method. Fichtner *et al.* (2017) present Mirmex, a complete tool for the processing of seismic
 91 ambient noise data that uses GPUs to compute the most demanding operations including
 92 direct implementations of the PCC and GNCC methods.

93 Conventional cross-correlations, and in particular GNCC, are computed directly only if
 94 a very small number of lag times are required; otherwise, they are computed much faster
 95 using the FFT. The FFT is very efficiently implemented on the CPU by the fastest Fourier
 96 transform in the west (FFTW) library (Frigo and Johnson, 2005) and can be further accel-
 97 erated on the GPU using, e.g., the compute unified device architecture (CUDA) platform.

98 Despite the advantages of PCC compared to the 1-bit cross-correlation, the application of
 99 this method to the large data volumes that current seismic applications demand is limited
 100 by its high computational cost.

101 In this article, we introduce two solutions to speed up the computation of PCC by a
 102 factor of 100 in conventional settings, so that PCC becomes only 2 to 4 times slower than
 103 conventional cross-correlation. We reduce operation complexity of the PCC when using
 104 modulus raised to the power of 2, and we present optimized CPU and GPU codes (see Data
 105 and Resources) to accelerate computations when using other powers, in particular, using
 106 the default power of 1 (PCC1) as recommended in Schimmel (1999) and Schimmel *et al.*
 107 (2011). These codes focus exclusively on the cross-correlation operation and allow for an
 108 easy integration with current processing flows.

109 Calculation of the phase cross-correlation

110 The PCC methods evaluates similarity between two analytic signals s_1 and s_2 at the lag
 111 time τ by subtracting the modulus of the sum and difference of their unitary phasors. To
 112 obtain these unitary phasors, we represent real time-series u as analytic signals s using the
 113 Hilbert transform $\mathcal{H}$,

$$s(t) = u(t) + \mathcal{H}\{u(t)\} = a(t) e^{i\theta(t)} \quad (1)$$

114 where a is envelope, θ is instantaneous phase and t is time. In this signal representation, we
 115 obtain unitary phasors containing the instantaneous phase information by dividing s by its
 116 modulus,

$$e^{i\theta(t)} = \frac{s(t)}{|s(t)|}. \quad (2)$$

117 Then, PCC uses the unitary phasors of the analytic signals $s_1(t) = a_1(t) e^{i\theta_1(t)}$ and $s_2(t) =$
 118 $a_2(t) e^{i\theta_2(t)}$ to compute the similarity between them.

119 The PCC of the discrete sequences $s_1[n]$ and $s_2[n]$ sampled at $t = t_0 + nT$, where n is

the sample number and T the sampling period, can be written as

$$c_{\text{pcc}}[m] = \frac{1}{N} \sum_{n=0}^{N-1} \left| \frac{e^{i\theta_1[n+m]} + e^{i\theta_2[n]}}{2} \right|^\nu - \left| \frac{e^{i\theta_1[n+m]} - e^{i\theta_2[n]}}{2} \right|^\nu \quad (3)$$

where m is the lag time number (from the lag time $\tau = mT$), ν is the power of PCC, and N is the number of samples correlated from t_0 to $t_0 + (N - 1)T$.

Analyzing eq. (3) in detail, we can see that if the signals are totally correlated, the first term in the sum equals to 1 and the second to 0, and thus $c_{\text{pcc}} = 1$; conversely, if they are totally anticorrelated, the first term is 0 and the second is 1, and thus $c_{\text{pcc}} = -1$; and if they are totally not correlated, the two terms are equal and $c_{\text{pcc}} = 0$. In this context, the power factor ν controls the sharpness of the transition from totally correlated and anticorrelated to not correlated values. Default values for ν are 1 and 2. Setting $\nu = 1$ increases absolute correlation values and setting $\nu > 1$, and particularly, $\nu = 2$, increases signal-to-noise ratio.

PCC with power factor 2

PCC as written in eq. (3) simplifies for $\nu = 2$ by developing the square modulus. The key operation to be implemented when computing the PCC with $\nu = 2$ (PCC2) at one lag time is

$$c = \sum_i |a_i + b_i|^2 - |a_i - b_i|^2 = \|a + b\|^2 - \|a - b\|^2 \quad (4)$$

where a_i and b_i are complex numbers, a and b are complex vectors and $\|\cdot\|$ denotes the Euclidean norm. The squared norm of the sum and difference of two elements in a complex Hilbert space can be developed as functions of inner products as

$$\begin{aligned} \|a + b\|^2 &= \langle a + b, a + b \rangle \\ &= \langle a, a \rangle + \langle a, b \rangle + \langle b, a \rangle + \langle b, b \rangle \end{aligned} \quad (5)$$

$$\begin{aligned} \|a - b\|^2 &= \langle a - b, a - b \rangle \\ &= \langle a, a \rangle - \langle a, b \rangle - \langle b, a \rangle + \langle b, b \rangle. \end{aligned} \quad (6)$$

137 Then, the difference of the two squared norms reduces to a single inner product,

$$\|a + b\|^2 - \|a - b\|^2 = 4 \operatorname{Re} \langle a, b \rangle, \quad (7)$$

138 which is the real part of the polarization identity for complex Hilbert spaces (Young, 1988).

139 When we apply this result to eq. (3), PCC2 simplifies to the real part of the complex

140 cross-correlation between the phase vectors,

$$c_{\text{pcc2}}[m] = \operatorname{Re} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} e^{i\theta_1[n+m]} e^{-i\theta_2[n]} \right\}, \quad (8)$$

141 or equivalently, using $s_1[n]$ and $s_2[n]$,

$$c_{\text{pcc2}}[m] = \operatorname{Re} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} \frac{s_1[n+m]}{|s_1[n+m]|} \frac{\overline{s_2[n]}}{|s_2[n]|} \right\} \quad (9)$$

142 where $\overline{s_2[n]}$ denotes the complex conjugate of $s_2[n]$.

143 Writing PCC2 as a conventional complex cross-correlation dramatically reduces its oper-

144 ation complexity and allows for further acceleration by the FFT. Overall, the computational

145 cost of PCC2 and GNCC are now comparable and the main differences are due to the re-

146 quired analytic signal representation for the PCC. Finally, it is important to note that eq.

147 (9) defines PCC2 if s_1 and s_2 are analytic signals and defines 1-bit GNCC if s_1 and s_2 are

148 real signals. Since the phase contains much more information about the signal than its sign,

149 it is reasonable to say that PCC can improve waveform coherence and increase convergence

150 compared to 1-bit GNCC in most applications.

151 Implementation

152 To our knowledge, the PCC equation with a power factor v different from 2 does not simplify

153 and the operation complexity cannot be reduced by algorithms such as the FFT; however, it

154 can be accelerated on the GPU. GPUs are optimized for computing the same operations on
 155 large datasets in parallel and CPUs for long sequential series of operations. Thus, GPUs are
 156 optimal to solve simple but computationally intensive operations on relatively small datasets
 157 such as PCC, where the wall-clock time reduction of the computations largely compensates
 158 for the time penalty of copying data to and from the GPU memory. Conversely, 1-bit GNCC
 159 and PCC2 are conventionally more efficiently implemented in the CPU using a few FFT
 160 operations than in the GPU, where the acceleration of the FFT computation is not worth
 161 the extra time required to move data between the CPU and GPU memories. To compute the
 162 FFT in the CPU we use the most recent iteration of the FFTW library (Frigo and Johnson,
 163 2005), a very efficient implementation that generally outperforms direct implementations of
 164 the cross-correlation. Only if the number of lag times is very small, less than about 1 per
 165 cent of the number of samples of the time series, are direct solutions typically faster.

166 The computation of PCC between two analytic sequences of N samples at M lag times
 167 for the default value of $v = 1$ involves $2NM$ sums and modulus of complex numbers and
 168 $(2N - 1)M$ sums of real numbers. The most demanding operation among them is by far the
 169 computation of the modulus of complex numbers, mainly due to the square root involved.
 170 This operation is common in many physical problems and efficiently computed on the GPU.

171 We develop CPU only and GPU enhanced codes of 1-bit GNCC and PCC (see Data and
 172 Resources) that compute all the correlations of a station pair at once. We use OpenMP to
 173 compute these correlations in parallel on the CPU, instead of parallelizing each correlation
 174 individually, to reduce scheduling overhead and improve scaling with the number of cores.
 175 To accelerate the PCC1 calculation when a GPU is available, we compute the analytic repre-
 176 sentation of the data sequences on the CPU using OpenMP and, in parallel, the subsequent
 177 operations on the GPU. To speed up computations of the actual correlations, eq. (3), we
 178 distribute the computations of one correlation across many GPU-thread blocks. Each thread
 179 in a block computes the correlation between 256 samples of the data sequences at one lag
 180 time using copies of the portions of the sequences needed on the small shared memory (much

181 faster and with lower latency than the global memory) available to each block. Moreover, we
 182 use 16 streams to compute several correlations in parallel, enough to fully exploit the cores
 183 available on modern GPUs. In this configuration, the GPU calculations limit wall-clock time
 184 and the computation of the analytic signals on the CPU reduces GPU usage and exploits
 185 resources that would be otherwise wasted.

186 Example

187 We illustrate the computational performance of 1-bit GNCC and PCC using continuous
 188 data on the vertical component. The pre-processing flow consists in removing the mean
 189 and trends, correcting the instrument response to produce ground velocity, band-pass filter-
 190 ing from 20 to 33 mHz (30 to 50 s period), dividing data into 1 day-long non-overlapping
 191 sequences, and rejecting sequences with high seismic activity either containing strong earth-
 192 quakes or having a mean energy much higher than the average. We specifically reject raw-
 193 data sequences with absolute maxima greater than the median plus 100 times the median
 194 absolute deviation (MAD). Then, we reject pre-processed sequences with a standard devia-
 195 tion greater than the median of the standard deviation of the sequences recorded by a given
 196 station plus 500 times its MAD. The test environment is a workstation equipped with a
 197 4-core 64-bit Intel Haswell CPU at 3.50 GHz (i5-4690K) with 16 GB of DDR3 at 1866 MHz,
 198 an Nvidia GTX 1060 graphics card, and an SSD disk to store the data.

199 Fig. 1 shows the wall-clock time cost of applying 1-bit GNCC, PCC2, and PCC1 with a
 200 maximum lag time of $\pm 12,000$ s to 649 day-long pairs of seismograms from the years 2016
 201 and 2017 recorded by the CAN and ECH GEOSCOPE stations and downsampled to 1,
 202 2, 4, 8 and 10 seconds per sample. We see that the time cost of 1-bit GNCC and PCC2
 203 have the same trend, mostly determined by FFT computations whose operation complexity
 204 is $\mathcal{O}(N \log_2 N)$, independent of the number of lag times M . PCC2 is about two times
 205 slower than 1-bit GNCC because the PCC2 implementation uses seven single-precision FFT

206 operations (four to compute the two analytic signals and three for the correlation) whereas 1-
 207 bit GNCC uses only three. In comparison, PCC1 on the CPU with an operation complexity
 208 of $\mathcal{O}(NM)$ is much slower and its computational cost increases faster. The GPU enhanced
 209 code of PCC1 substantially reduces the wall-clock time of the CPU-only code and makes
 210 PCC1 only about four times slower than 1-bit GNCC in conventional settings. For example,
 211 PCC1 takes 71.65 s on the CPU and 0.70 s on the GPU for the sequence length of 21,600
 212 samples per day, which is 4.17 times slower than 1-bit GNCC and 1.79 times slower than
 213 PCC2, and 1.9 s on a single NVIDIA Tesla K20 GPU from the S-CAPAD IPGP cluster.

214 In the next example, we have selected data from the years 1990 to 2017 recorded by 299
 215 broadband seismic stations distributed worldwide as shown in Fig. 2. In particular, we have
 216 selected all stations from the global and regional networks of GEOFON (GE), GEOSCOPE
 217 (G), GSN (IU,II) and New China Digital Seismograph Network (IC). In addition, we have
 218 included some stations from Geoscience Australia (AU), the Canadian National Seismograph
 219 Network (CN), the Danish Seismological Network (DK), the Pacific21 (PS) network, and the
 220 Canfranc Underground Laboratory (LC) to improve global Earth coverage.

221 We have computed 1-bit GNCC, PCC2, and PCC1 of day-long sequences downsampled
 222 to sampling periods of 4 seconds per sample. Then, we have rejected correlations with
 223 an anomalous standard deviation and stack the rest linearly. Under these settings, the
 224 faster algorithms would have spent most of the time on reading and writing operations. To
 225 accelerate these operations, we have gathered all data from a given station in one file and
 226 we have used a ramdisk for temporal results such as the daily correlations. Using the above
 227 workstation, the correlation and stack operations require 939 minutes for 1-bit GNCC, 1,291
 228 minutes for PCC2, and 2,380 for PCC1 on the GPU.

229 Fig. 3(a) shows 110 million day-long interstation correlations using PCC2 from 39,656
 230 station pairs band-pass filtered from 20 to 33 mHz as a function of distance stacked on bins
 231 of 0.25 degrees using only data from days with low seismic activity. The large amount of
 232 day-long correlations stacked on each bin (Fig. 3b) is sufficient for the above correlation

233 methods to converge to the empirical Green’s function. In applications where few data
234 is available, PCC usually contributes to improve signal quality and increases convergence
235 (Schimmel *et al.*, 2011). The amplitude of the signals extracted is meaningful, in spite of the
236 aggressive amplitude normalization of 1-bit GNCC or PCC, but strongly influenced by the
237 distribution of noise sources (e.g., Snieder, 2004). Hence, to extract accurate information
238 from amplitudes it is important to design processing flows that minimize this bias as, for
239 example, Bowden *et al.* (2018) show in the construction of surface wave attenuation maps
240 across the US.

241 The strongest signal we observe in Fig. 3 is the Rayleigh wave starting at the bottom-left
242 corner and progressively vanishing at larger distance due to heterogeneities in the crust and
243 the top upper mantle. Much weaker, we see a great variety of body waves interacting with
244 the mantle and the core such as the P, PcP, S, ScS, PKP, PKIKP (I), or PKIKPPKIKP
245 (I2), and of noncausal phases (Phạm *et al.*, 2018) such as PKP-PcP preceding the direct P
246 wave arrival in the range $120 - 140^\circ$ by about 4 min. The number of station pairs per bin
247 from 5 to 162 degrees varies from about 20 to 100 pairs; however, it reduces to a few pairs
248 close to 180 degrees. At these interstation distances some of these signals may be interpreted
249 as spurious arrivals from dominant sources located far from the great circle between some
250 station pairs (Retailleau *et al.*, 2017).

251 The signals we observe in this figure are produced by a mix of ambient noise sources
252 and coda waves from moderate and smaller earthquakes, as can be seen from the anomalous
253 high amplitude of the ScS, PKP, and PKIKP waves. Removing anomalously correlations
254 and sequences with high seismic activity have removed contributions from large earthquakes
255 whose poor distribution in azimuth may introduce bias in, e.g., velocity dispersion or arrival
256 time observations. More effective processing flows for removing coda waves can be found
257 in Nishida (2013) and Boué *et al.* (2014). Phạm *et al.* (2018) observe some of the signals
258 we show in Fig. 3 using coda from large earthquakes instead and explain the origin and
259 generation mechanism of causal and noncausal phases. Further, in Fig. 3, we observe signals

260 with low apparent velocity, such as P, S, PS and PPS, mainly produced by ambient noise
261 sources.

262 Conclusions

263 We have introduced approaches that enable the application of PCC on large data volumes
264 using conventional workstations. The stand-alone code implementing PCC and 1-bit GNCC
265 is designed to be easily integrated with current processing flows, whether it is executed
266 directly or the main functions implementing each method are called from other codes.

267 We have shown that PCC simplifies to a conventional cross-correlation of unitary phasors
268 when using the power factor 2. This cross-correlation equation is fully analogous to 1-bit
269 GNCC with the single difference that 1-bit GNCC considers the sign of the real signals and
270 PCC2 the phase factor of analytic signals. Since phase carries more information than sign,
271 most applications using 1-bit GNCC should see a gain in waveform coherence and increased
272 convergence when applying PCC2.

273 By writing PCC2 as a complex cross-correlation, we have achieved a lower operation
274 complexity and, in conventional settings, about 100 times faster code than direct imple-
275 mentations. This reduces the computational cost of PCC2 to about two times that of the
276 1-bit GNCC method using only the CPU. When using other powers, GPU acceleration has
277 allowed us to apply PCC1 to the large data volume needed to extract body wave signals
278 from interstation correlations using a conventional workstation and a low-cost GPU. In gen-
279 eral, these computational cost reductions are important to efficiently find and tune the best
280 adapted methods to a given problem and, especially, to process the large data volumes
281 that modern seismic ambient noise imaging and monitoring applications demand with phase
282 cross-correlation.

283 Data and Resources

284 The data used in this study were obtained from the Incorporated Research Institutions of
285 Seismology (IRIS) and the European Integrated Data Archive (EIDA). The CPU and GPU
286 accelerated versions of code implementing 1-bit GNCC and PCC are open source under
287 the LGPL v3 license and available at <https://github.com/sergiventosa/FastPCC>. The
288 fastest Fourier transform in the west (FFTW) is available from <http://www.fftw3.org> and
289 the CUDA documentation and installation guides from [https://docs.nvidia.com/cuda/](https://docs.nvidia.com/cuda/index.html)
290 [index.html](https://docs.nvidia.com/cuda/index.html).

291 Acknowledgments

292 We thank ~~to the editor~~ the editors and the anonymous reviewers for improving the manuscript.
293 This work was supported by the MISTERIOS project CGL2013-48601-C2-1-R and the ANR
294 MIMOSA project ANR-14-CE01-0012. We also acknowledge support from discussions within
295 TIDES COST Actions ES1401. Numerical computations were partly performed on the S-
296 CAPAD platform, IPGP, France.

297 References

- 298 Baig, A., M. Campillo, and F. Brenguier (2009). Denoising seismic noise cross correlations,
299 *J. Geophys. Res.*, **114**, no. B8, 1–12, doi:10.1029/2008JB006085.
- 300 Becker, G., and B. Knapmeyer-Endrun (2018). Crustal thickness across the Trans-European
301 Suture Zone from ambient noise autocorrelations, *Geophys. J. Int.*, **212**, no. 2, 1237–1254,
302 doi:10.1093/gji/ggx485.
- 303 Bensen, G. D., M. H. Ritzwoller, M. P. Barmin, A. L. Levshin, F. Lin, M. P. Moschetti,
304 N. M. Shapiro, and Y. Yang (2007). Processing seismic ambient noise data to obtain

305 reliable broad-band surface wave dispersion measurements, *Geophys. J. Int.*, **169**, no. 3,
306 1239–1260, doi:10.1111/j.1365-246x.2007.03374.x.

307 Boué, P., P. Poli, M. Campillo, and P. Roux (2014). Reverberations, coda waves and ambient
308 noise: Correlations at the global scale and retrieval of the deep phases, *Earth Planet. Sci.*
309 *Lett.*, **391**, 137–145, doi:10.1016/j.epsl.2014.01.047.

310 Bowden, D. C., V. C. Tsai, and F.-C. Lin (2018). Amplification and attenuation across
311 USArray using ambient noise wavefront tracking, *J. Geophys. Res.*, **122**, no. 12, 10,086–
312 10,101, doi:10.1002/2017JB014804

313 Cheng, F., J. Xia, Y. Xu, Z. Xu, and Y. Pan (2015). A new passive seismic method based
314 on seismic interferometry and multichannel analysis of surface waves, *J. Appl. Geophys.*,
315 **117**, 126–135, doi:10.1016/j.jappgeo.2015.04.005.

316 Cupillard, P., L. Stehly, and B. Romanowicz (2011). The one-bit noise correlation: a theory
317 based on the concepts of coherent and incoherent noise, *Geophys. J. Int.*, **184**, no. 3,
318 1397–1414, doi:10.1111/j.1365-246x.2010.04923.x.

319 D’Hour, V., M. Schimmel, A. F. Nascimento, J. M. Ferreira, and H. C. Lima Neto
320 (2016). Detection of subtle hydromechanical medium changes caused by a small-magnitude
321 earthquake swarm in NE brazil, *Pure Appl. Geophys.*, **173**, no. 4, 1097–1113, doi:
322 10.1007/s00024-015-1156-0.

323 Fichtner, A., L. Ermert, and A. Gokhberg (2017). Seismic noise correlation on heterogeneous
324 supercomputers, *Seismol. Res. Lett.*, **88**, no. 4, 1141–1145, doi:10.1785/0220170043.

325 Frigo, M., and S. G. Johnson (2005). The design and implementation of FFTW3, *Proc. of*
326 *the IEEE*, **93**, no. 2, 216–231, doi:10.1109/JPROC.2004.840301, special issue on “Program
327 Generation, Optimization, and Platform Adaptation”.

328 Hadziioannou, C., E. Larose, A. Baig, P. Roux, and M. Campillo (2011). Improving temporal
 329 resolution in ambient noise monitoring of seismic wave speed, *J. Geophys. Res.*, **116**, no.
 330 B7, 1–10, doi:10.1029/2011JB008200.

331 Haned, A., E. Stutzmann, M. Schimmel, S. Kiselev, A. Davaille, and A. Yelles-Chaouche
 332 (2016). Global tomography using seismic hum, *Geophys. J. Int.*, **204**, no. 2, 1222–1236,
 333 doi:10.1093/gji/ggv516.

334 Krischer, L., T. Megies, R. Barsch, M. Beyreuther, T. Lecocq, C. Caudron, and J. Wasser-
 335 mann (2015). ObsPy: a bridge for seismology into the scientific Python ecosystem, *Com-
 336 put. Sci. Disc.*, **8**, no. 1, 014,003, doi:10.1088/1749-4699/8/1/014003.

337 Lecocq, T., C. Caudron, and F. Brenguier (2014). MSNoise, a Python package for monitoring
 338 seismic velocity changes using ambient seismic noise, *Seismol. Res. Lett.*, **85**, no. 3, 715–
 339 726, doi:10.1785/0220130073.

340 Moreau, L., L. Stehly, P. Boué, Y. Lu, E. Larose, and M. Campillo (2017). Improving ambient
 341 noise correlation functions with an SVD-based Wiener filter, *Geophys. J. Int.*, **211**, no. 1,
 342 418–426, doi:10.1093/gji/ggx306.

343 Nishida, K. (2013). Global propagation of body waves revealed by cross-correlation analysis
 344 of seismic hum, *Geophys. Res. Lett.*, **40**, no. 9, 1691–1696, doi:10.1002/grl.50269.

345 Obermann, A., M. Lupi, A. Mordret, S. S. Jakobsdóttir, and S. A. Miller (2016). 3D-ambient
 346 noise Rayleigh wave tomography of Snæfellsjökull volcano, iceland, *J. Volcanol. Geoth.
 347 Res.*, **317**, 42–52, doi:10.1016/j.jvolgeores.2016.02.013.

348 Pham, T.-S., H. Tkalčić, M. Sambridge, and B. L. N. Kennett (2018). Earth’s correlation
 349 wavefield: Late coda correlation, *Geophys. Res. Lett.*, **45**, no. 7, 3035–3042, doi:10.1002/
 350 2018GL077244.

351 Retailleau, L., P. Boué, L. Stehly, and M. Campillo (2017). Locating microseism sources
 352 using spurious arrivals in intercontinental noise correlations, *J. Geophys. Res.*, **122**, no.
 353 10, 8107–8120, doi:10.1002/2017JB014593.

354 Romero, P., and M. Schimmel (2018). Mapping the basement of the Ebro basin in Spain
 355 with seismic ambient noise autocorrelations, *J. Geophys. Res.*, **123**, no. 6, 5052–5067,
 356 doi:10.1029/2018JB015498.

357 Sánchez-Pastor, P., A. Obermann, and M. Schimmel (2018). Detecting and locating precur-
 358 sory signals during the 2011 El Hierro, Canary Islands, submarine eruption, *Geophys. Res.*
 359 *Lett.*, **45**, no. 19, 10,288–10,297, doi:10.1029/2018GL079550.

360 Schimmel, M. (1999). Phase cross-correlations: Design, comparisons, and applications, *Bull.*
 361 *Seismol. Soc. Am.*, **89**, no. 5, 1366–1378.

362 Schimmel, M., E. Stutzmann, and J. Gallart (2011). Using instantaneous phase coherence
 363 for signal extraction from ambient noise data at a local to a global scale, *Geophys. J. Int.*,
 364 **184**, no. 1, 494–506, doi:10.1111/j.1365-246X.2010.04861.x.

365 Schimmel, M., E. Stutzmann, and S. Ventosa (2018). Low-frequency ambient noise auto-
 366 correlations: waveforms and normal modes, *Seismol. Res. Lett.*, **89**, no. 4, 1488–1496,
 367 doi:10.1785/0220180027.

368 Schippkus, S., D. Zigone, G. Bokelmann, and the AlpArray Working Group (2018). Ambient-
 369 noise tomography of the wider Vienna basin region, *Geophys. J. Int.*, **215**, no. 1, 102–117,
 370 doi:10.1093/gji/ggy259.

371 Snieder, R. (2004). Extracting the Green’s function from the correlation of coda waves: A
 372 derivation based on stationary phase, *Phys. Rev. E*, **69**, 046,610, doi:10.1103/PhysRevE.
 373 69.046610.

374 Takano, T., T. Nishimura, and H. Nakahara (2017). Seismic velocity changes concentrated
 375 at the shallow structure as inferred from correlation analyses of ambient noise during
 376 volcano deformation at Izu-Oshima, Japan, *J. Geophys. Res.*, **122**, no. 8, 6721–6736, doi:
 377 10.1002/2017JB014340.

378 Taylor, G., S. Rost, and G. Houseman (2016). Crustal imaging across the North Anatolian
 379 Fault Zone from the auto-correlation of ambient seismic noise, *Geophys. Res. Lett.*, **43**,
 380 no. 6, 2502–2509, doi:10.1002/2016GL067715.

381 Tkalčić, H., and T.-S. Phạm (2018). Shear properties of Earth’s inner core constrained by
 382 a detection of J waves in global correlation wavefield, *Science*, **362**, no. 6412, 329–332,
 383 doi:10.1126/science.aau7649.

384 Ventosa, S., M. Schimmel, and E. Stutzmann (2017). Extracting surface waves, hum and
 385 normal modes: time-scale phase-weighted stack and beyond, *Geophys. J. Int.*, **211**, no. 1,
 386 30–44, doi:10.1093/gji/ggx284.

387 Young, N. (1988). *An introduction to Hilbert space*, Cambridge University Press.

388 Zeng, X., and C. Thurber (2016). A graphics processing unit implementation for time-
 389 frequency phase-weighted stacking, *Seismol. Res. Lett.*, **87**, no. 2A, 358–362, doi:10.1785/
 390 0220150192.

391 List of Figures

392	1	Computational cost of cross-correlating 649 days of data from the CAN and	
393		ECH GEOSCOPE stations using 1-bit CC, PCC2 and PCC1 with a maximum	
394		correlation lag of $\pm 12,000$ s.	19
395	2	Geographic distribution of the stations (red triangles) used in this test. . . .	20
396	3	(a) 110 million day-long interstation correlations from 39,656 station pairs	
397		sorted by distance and stacked in bins of 0.25 degrees in the period band of	
398		30 to 50 s using PCC2 and linear stacking on days with low seismic activity.	
399		(b) Day-long correlations stacked per bin.	21

Figure 1: Computational cost of cross-correlating 649 days of data from the CAN and ECH GEOSCOPE stations using 1-bit CC, PCC2 and PCC1 with a maximum correlation lag of $\pm 12,000$ s.

Figure 2: Geographic distribution of the stations (red triangles) used in this test.

Figure 3: (a) 110 million day-long interstation correlations from 39,656 station pairs sorted by distance and stacked in bins of 0.25 degrees in the period band of 30 to 50 s using PCC2 and linear stacking on days with low seismic activity. (b) Day-long correlations stacked per bin.

Towards the processing of large data volumes with phase cross-correlation

Sergi Ventosa*, Martin Schimmel*, Eleonore Stutzmann†

Abstract

Interstation correlation is the basic operation in seismic noise and coda wave interferometry for signal extraction in imaging and monitoring applications. Conventional cross-correlations evaluate the similarity between two signals along lag time and they are efficiently computed by the fast Fourier transform (FFT), valuable to manage the large data volumes that ambient noise applications demand. The phase cross-correlation (PCC) method contributes to increase convergence, a key issue in seismic ambient noise imaging and monitoring; however, it is much more computationally demanding. PCC evaluates similarity by subtracting the modulus of the sum and difference of the instantaneous phase of two signals. We introduce solutions to dramatically reduce the high computational cost of PCC. We show that PCC can be rewritten as a complex cross-correlation and computed by the FFT when the moduli are raised to the power of 2, and we demonstrate PCC can improve waveform coherence and increase convergence compared to the default processing flow of 1-bit amplitude normalization and standard cross-correlation. Moreover, we develop a graphics processing unit implementation to accelerate computations when using powers other than 2 and particularly

*Institute of Earth Sciences Jaume Almera - CSIC, E-08028 Barcelona, Spain.

†Institut de Physique du Globe de Paris, Sorbonne Paris Cité, Univ Paris Diderot, UMR 7154 CNRS, F-75005 Paris, France.

19 when using the power of 1. Finally, we extract Rayleigh and body wave signals from many
20 years of data from seismic stations distributed worldwide using PCC without a significant
21 increase in computational cost compared to conventional cross-correlation.

22 Introduction

23 Cross-correlation is a ubiquitous method in seismic data processing important for picking,
24 empirical Green’s function extraction, ambient noise monitoring, waveform comparison, and
25 signal, event and pattern detection or identification among others. Phase cross-correlation
26 (PCC) was conceived to better detect weak P-to-S conversions at the upper mantle disconti-
27 nities (Schimmel, 1999). In this work, PCC is introduced to estimate travel-time differences
28 by searching for the lag time that maximize an amplitude-unbiased misfit function between
29 the pilot and converted waveforms, their phase coherency. Phase coherency is useful for
30 detecting weak signals in the coda of stronger waves because it effectively considers a larger
31 portion of the waveforms than conventional correlation-based methods that consider mainly
32 the most energetic components. Later on, PCC was found useful to improve waveform coher-
33 ence and convergence of interstation correlations thanks to its amplitude-unbiased property
34 (Schimmel *et al.*, 2011). Essentially, PCC can be useful for applications demanding more
35 information on the waveform.

36 Interstation correlations of seismic ambient noise and scattered coda waves complement
37 direct arrivals from earthquakes in surface-wave tomography, monitoring, and Earth’s deep
38 interior studies in regions otherwise poorly illuminated. Convergence to the empirical Green’s
39 function or a robust seismic signal extraction increases compared to the direct correlation of
40 raw seismic records through customized processing flows that improve the balance of noise
41 sources and reduce the influence of anomalous signals. These processing flows compute many
42 correlations on relatively short data sequences. The pre-processing of the raw data may in-
43 clude instrument response correction, anomalous signal rejection, spectral whitening, and

time-domain normalization (Bensen *et al.*, 2007), and once the correlations are computed, weighted stacking and denoising (Baig *et al.*, 2009; Schimmel *et al.*, 2011; Cheng *et al.*, 2015; Moreau *et al.*, 2017). Finally, subsequent measurements of group and phase velocity dispersion are done in surface-wave tomography applications (e.g., Haned *et al.*, 2016; Schippkus *et al.*, 2018; Obermann *et al.*, 2016), of velocity and structure variations in monitoring (e.g., Hadziioannou *et al.*, 2011; D’Hour *et al.*, 2016; Takano *et al.*, 2017; Sánchez-Pastor *et al.*, 2018), or of travel times of body waves from ambient noise or earthquake coda in Earth’s deep interior studies (e.g., Nishida, 2013; Boué *et al.*, 2014; Phạm *et al.*, 2018; Tkalčić and Phạm, 2018).

Spectral whitening or band-pass filtering to the frequency band of interest followed by 1-bit amplitude normalization and geometrical-normalized cross-correlation (GNCC) is the processing flow most often used, where GNCC is the conventional cross-correlation divided by the standard deviation of the signals. The 1-bit cross-correlation method applies correlation on the sign of the input sequences (e.g., Bensen *et al.*, 2007; Cupillard *et al.*, 2011). This instantaneous amplitude normalization reduces the influence of poorly-distributed high-energy features such as earthquakes but it distorts waveforms and slows down convergence (Schimmel *et al.*, 2018). The PCC method measures similarity between two analytic signals of unitary amplitude, i.e., two instantaneous phases. Schimmel *et al.* (2011) show that PCC is a less aggressive alternative to 1-bit cross-correlation that may help to improve waveform coherence and increase convergence. This has been confirmed, for example, in D’Hour *et al.* (2016) for monitoring of hydromechanical changes, in Taylor *et al.* (2016) and Becker and Knapmeyer-Endrun (2018) for retrieving P wave Moho reflections from noise autocorrelations, in Haned *et al.* (2016) for extracting group-velocity observations for global tomography using seismic hum, in Schimmel *et al.* (2018) for extracting Rayleigh wave group velocities using phase autocorrelation, and in Romero and Schimmel (2018) for mapping the basement of the Ebro Basin in Spain using shallow subsurface reflections.

The computational cost of the signal processing methods conventionally used in the pre-

71 processing of the raw time series is small compared to the correlation, weighted stacking,
 72 and denoising operations subsequently applied. This is mainly, because pre-processing is
 73 done per station and correlation and denoising operations are done per station pair. Ob-
 74 sPy (Krischer *et al.*, 2015) and MSNoise (Lecocq *et al.*, 2014) are open-source packages for
 75 Python that are excellent for proof-of-concept studies and often sufficient for pre-processing.
 76 Compiled programming language solutions using the central processing unit (CPU) or the
 77 graphics processing unit (GPU) are however necessary in order to apply the most demand-
 78 ing operations to large data volumes in a reasonable amount of time. Several tools have
 79 been introduced recently for improving the computational efficiency of key operations along
 80 interstation correlation processing flows. Zeng and Thurber (2016) present a GPU imple-
 81 mentation of the time-frequency phase-weighted stack (tf-PWS; Schimmel *et al.*, 2011) which
 82 is up to 20 times faster than equivalent CPU implementations using the FFT. Ventosa *et al.*
 83 (2017) reduce computational cost of tf-PWS by introducing the time-scale phase-weighted
 84 stack (ts-PWS) that uses the wavelet transform for the time-frequency representation, and
 85 propose unbiased phase coherence and two-stage ts-PWS to improve convergence speed and
 86 quality of the extracted signals. The unbiased phase coherence estimator contributes to
 87 increase non-coherent noise attenuation because it equals to 0 if the signals are totally inco-
 88 herent. Whereas the two-stage ts-PWS reduces signal attenuation by stacking linearly the
 89 sequences in a few groups to improve signal-to-noise ratio before applying the actual ts-PWS
 90 method. Fichtner *et al.* (2017) present Mirmex, a complete tool for the processing of seismic
 91 ambient noise data that uses GPUs to compute the most demanding operations including
 92 direct implementations of the PCC and GNCC methods.

93 Conventional cross-correlations, and in particular GNCC, are computed directly only if
 94 a very small number of lag times are required; otherwise, they are computed much faster
 95 using the FFT. The FFT is very efficiently implemented on the CPU by the fastest Fourier
 96 transform in the west (FFTW) library (Frigo and Johnson, 2005) and can be further accel-
 97 erated on the GPU using, e.g., the compute unified device architecture (CUDA) platform.

98 Despite the advantages of PCC compared to the 1-bit cross-correlation, the application of
 99 this method to the large data volumes that current seismic applications demand is limited
 100 by its high computational cost.

101 In this article, we introduce two solutions to speed up the computation of PCC by a
 102 factor of 100 in conventional settings, so that PCC becomes only 2 to 4 times slower than
 103 conventional cross-correlation. We reduce operation complexity of the PCC when using
 104 modulus raised to the power of 2, and we present optimized CPU and GPU codes (see Data
 105 and Resources) to accelerate computations when using other powers, in particular, using
 106 the default power of 1 (PCC1) as recommended in Schimmel (1999) and Schimmel *et al.*
 107 (2011). These codes focus exclusively on the cross-correlation operation and allow for an
 108 easy integration with current processing flows.

109 Calculation of the phase cross-correlation

110 The PCC methods evaluates similarity between two analytic signals s_1 and s_2 at the lag
 111 time τ by subtracting the modulus of the sum and difference of their unitary phasors. To
 112 obtain these unitary phasors, we represent real time-series u as analytic signals s using the
 113 Hilbert transform $\mathcal{H}$,

$$s(t) = u(t) + \mathcal{H}\{u(t)\} = a(t) e^{i\theta(t)} \quad (1)$$

114 where a is envelope, θ is instantaneous phase and t is time. In this signal representation, we
 115 obtain unitary phasors containing the instantaneous phase information by dividing s by its
 116 modulus,

$$e^{i\theta(t)} = \frac{s(t)}{|s(t)|}. \quad (2)$$

117 Then, PCC uses the unitary phasors of the analytic signals $s_1(t) = a_1(t) e^{i\theta_1(t)}$ and $s_2(t) =$
 118 $a_2(t) e^{i\theta_2(t)}$ to compute the similarity between them.

119 The PCC of the discrete sequences $s_1[n]$ and $s_2[n]$ sampled at $t = t_0 + nT$, where n is

the sample number and T the sampling period, can be written as

$$c_{\text{pcc}}[m] = \frac{1}{N} \sum_{n=0}^{N-1} \left| \frac{e^{i\theta_1[n+m]} + e^{i\theta_2[n]}}{2} \right|^\nu - \left| \frac{e^{i\theta_1[n+m]} - e^{i\theta_2[n]}}{2} \right|^\nu \quad (3)$$

where m is the lag time number (from the lag time $\tau = mT$), ν is the power of PCC, and N is the number of samples correlated from t_0 to $t_0 + (N - 1)T$.

Analyzing eq. (3) in detail, we can see that if the signals are totally correlated, the first term in the sum equals to 1 and the second to 0, and thus $c_{\text{pcc}} = 1$; conversely, if they are totally anticorrelated, the first term is 0 and the second is 1, and thus $c_{\text{pcc}} = -1$; and if they are totally not correlated, the two terms are equal and $c_{\text{pcc}} = 0$. In this context, the power factor ν controls the sharpness of the transition from totally correlated and anticorrelated to not correlated values. Default values for ν are 1 and 2. Setting $\nu = 1$ increases absolute correlation values and setting $\nu > 1$, and particularly, $\nu = 2$, increases signal-to-noise ratio.

PCC with power factor 2

PCC as written in eq. (3) simplifies for $\nu = 2$ by developing the square modulus. The key operation to be implemented when computing the PCC with $\nu = 2$ (PCC2) at one lag time is

$$c = \sum_i |a_i + b_i|^2 - |a_i - b_i|^2 = \|a + b\|^2 - \|a - b\|^2 \quad (4)$$

where a_i and b_i are complex numbers, a and b are complex vectors and $\|\cdot\|$ denotes the Euclidean norm. The squared norm of the sum and difference of two elements in a complex Hilbert space can be developed as functions of inner products as

$$\begin{aligned} \|a + b\|^2 &= \langle a + b, a + b \rangle \\ &= \langle a, a \rangle + \langle a, b \rangle + \langle b, a \rangle + \langle b, b \rangle \end{aligned} \quad (5)$$

$$\begin{aligned} \|a - b\|^2 &= \langle a - b, a - b \rangle \\ &= \langle a, a \rangle - \langle a, b \rangle - \langle b, a \rangle + \langle b, b \rangle. \end{aligned} \quad (6)$$

137 Then, the difference of the two squared norms reduces to a single inner product,

$$\|a + b\|^2 - \|a - b\|^2 = 4 \operatorname{Re} \langle a, b \rangle, \quad (7)$$

138 which is the real part of the polarization identity for complex Hilbert spaces (Young, 1988).

139 When we apply this result to eq. (3), PCC2 simplifies to the real part of the complex

140 cross-correlation between the phase vectors,

$$c_{\text{pcc2}}[m] = \operatorname{Re} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} e^{i\theta_1[n+m]} e^{-i\theta_2[n]} \right\}, \quad (8)$$

141 or equivalently, using $s_1[n]$ and $s_2[n]$,

$$c_{\text{pcc2}}[m] = \operatorname{Re} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} \frac{s_1[n+m]}{|s_1[n+m]|} \frac{\overline{s_2[n]}}{|s_2[n]|} \right\} \quad (9)$$

142 where $\overline{s_2[n]}$ denotes the complex conjugate of $s_2[n]$.

143 Writing PCC2 as a conventional complex cross-correlation dramatically reduces its oper-
 144 ation complexity and allows for further acceleration by the FFT. Overall, the computational
 145 cost of PCC2 and GNCC are now comparable and the main differences are due to the re-
 146 quired analytic signal representation for the PCC. Finally, it is important to note that eq.
 147 (9) defines PCC2 if s_1 and s_2 are analytic signals and defines 1-bit GNCC if s_1 and s_2 are
 148 real signals. Since the phase contains much more information about the signal than its sign,
 149 it is reasonable to say that PCC can improve waveform coherence and increase convergence
 150 compared to 1-bit GNCC in most applications.

151 Implementation

152 To our knowledge, the PCC equation with a power factor v different from 2 does not simplify
 153 and the operation complexity cannot be reduced by algorithms such as the FFT; however, it

154 can be accelerated on the GPU. GPUs are optimized for computing the same operations on
 155 large datasets in parallel and CPUs for long sequential series of operations. Thus, GPUs are
 156 optimal to solve simple but computationally intensive operations on relatively small datasets
 157 such as PCC, where the wall-clock time reduction of the computations largely compensates
 158 for the time penalty of copying data to and from the GPU memory. Conversely, 1-bit GNCC
 159 and PCC2 are conventionally more efficiently implemented in the CPU using a few FFT
 160 operations than in the GPU, where the acceleration of the FFT computation is not worth
 161 the extra time required to move data between the CPU and GPU memories. To compute the
 162 FFT in the CPU we use the most recent iteration of the FFTW library (Frigo and Johnson,
 163 2005), a very efficient implementation that generally outperforms direct implementations of
 164 the cross-correlation. Only if the number of lag times is very small, less than about 1 per
 165 cent of the number of samples of the time series, are direct solutions typically faster.

166 The computation of PCC between two analytic sequences of N samples at M lag times
 167 for the default value of $v = 1$ involves $2NM$ sums and modulus of complex numbers and
 168 $(2N - 1)M$ sums of real numbers. The most demanding operation among them is by far the
 169 computation of the modulus of complex numbers, mainly due to the square root involved.
 170 This operation is common in many physical problems and efficiently computed on the GPU.

171 We develop CPU only and GPU enhanced codes of 1-bit GNCC and PCC (see Data and
 172 Resources) that compute all the correlations of a station pair at once. We use OpenMP to
 173 compute these correlations in parallel on the CPU, instead of parallelizing each correlation
 174 individually, to reduce scheduling overhead and improve scaling with the number of cores.
 175 To accelerate the PCC1 calculation when a GPU is available, we compute the analytic repre-
 176 sentation of the data sequences on the CPU using OpenMP and, in parallel, the subsequent
 177 operations on the GPU. To speed up computations of the actual correlations, eq. (3), we
 178 distribute the computations of one correlation across many GPU-thread blocks. Each thread
 179 in a block computes the correlation between 256 samples of the data sequences at one lag
 180 time using copies of the portions of the sequences needed on the small shared memory (much

181 faster and with lower latency than the global memory) available to each block. Moreover, we
 182 use 16 streams to compute several correlations in parallel, enough to fully exploit the cores
 183 available on modern GPUs. In this configuration, the GPU calculations limit wall-clock time
 184 and the computation of the analytic signals on the CPU reduces GPU usage and exploits
 185 resources that would be otherwise wasted.

186 Example

187 We illustrate the computational performance of 1-bit GNCC and PCC using continuous
 188 data on the vertical component. The pre-processing flow consists in removing the mean
 189 and trends, correcting the instrument response to produce ground velocity, band-pass filter-
 190 ing from 20 to 33 mHz (30 to 50 s period), dividing data into 1 day-long non-overlapping
 191 sequences, and rejecting sequences with high seismic activity either containing strong earth-
 192 quakes or having a mean energy much higher than the average. We specifically reject raw-
 193 data sequences with absolute maxima greater than the median plus 100 times the median
 194 absolute deviation (MAD). Then, we reject pre-processed sequences with a standard devia-
 195 tion greater than the median of the standard deviation of the sequences recorded by a given
 196 station plus 500 times its MAD. The test environment is a workstation equipped with a
 197 4-core 64-bit Intel Haswell CPU at 3.50 GHz (i5-4690K) with 16 GB of DDR3 at 1866 MHz,
 198 an Nvidia GTX 1060 graphics card, and an SSD disk to store the data.

199 Fig. 1 shows the wall-clock time cost of applying 1-bit GNCC, PCC2, and PCC1 with a
 200 maximum lag time of $\pm 12,000$ s to 649 day-long pairs of seismograms from the years 2016
 201 and 2017 recorded by the CAN and ECH GEOSCOPE stations and downsampled to 1,
 202 2, 4, 8 and 10 seconds per sample. We see that the time cost of 1-bit GNCC and PCC2
 203 have the same trend, mostly determined by FFT computations whose operation complexity
 204 is $\mathcal{O}(N \log_2 N)$, independent of the number of lag times M . PCC2 is about two times
 205 slower than 1-bit GNCC because the PCC2 implementation uses seven single-precision FFT

operations (four to compute the two analytic signals and three for the correlation) whereas 1-bit GNCC uses only three. In comparison, PCC1 on the CPU with an operation complexity of $\mathcal{O}(NM)$ is much slower and its computational cost increases faster. The GPU enhanced code of PCC1 substantially reduces the wall-clock time of the CPU-only code and makes PCC1 only about four times slower than 1-bit GNCC in conventional settings. For example, PCC1 takes 71.65 s on the CPU and 0.70 s on the GPU for the sequence length of 21,600 samples per day, which is 4.17 times slower than 1-bit GNCC and 1.79 times slower than PCC2, and 1.9 s on a single NVIDIA Tesla K20 GPU from the S-CAPAD IPGP cluster.

In the next example, we have selected data from the years 1990 to 2017 recorded by 299 broadband seismic stations distributed worldwide as shown in Fig. 2. In particular, we have selected all stations from the global and regional networks of GEOFON (GE), GEOSCOPE (G), GSN (IU,II) and New China Digital Seismograph Network (IC). In addition, we have included some stations from Geoscience Australia (AU), the Canadian National Seismograph Network (CN), the Danish Seismological Network (DK), the Pacific21 (PS) network, and the Canfranc Underground Laboratory (LC) to improve global Earth coverage.

We have computed 1-bit GNCC, PCC2, and PCC1 of day-long sequences downsampled to sampling periods of 4 seconds per sample. Then, we have rejected correlations with an anomalous standard deviation and stack the rest linearly. Under these settings, the faster algorithms would have spent most of the time on reading and writing operations. To accelerate these operations, we have gathered all data from a given station in one file and we have used a ramdisk for temporal results such as the daily correlations. Using the above workstation, the correlation and stack operations require 939 minutes for 1-bit GNCC, 1,291 minutes for PCC2, and 2,380 for PCC1 on the GPU.

Fig. 3(a) shows 110 million day-long interstation correlations using PCC2 from 39,656 station pairs band-pass filtered from 20 to 33 mHz as a function of distance stacked on bins of 0.25 degrees using only data from days with low seismic activity. The large amount of day-long correlations stacked on each bin (Fig. 3b) is sufficient for the above correlation

233 methods to converge to the empirical Green’s function. In applications where few data
 234 is available, PCC usually contributes to improve signal quality and increases convergence
 235 (Schimmel *et al.*, 2011). The amplitude of the signals extracted is meaningful, in spite of the
 236 aggressive amplitude normalization of 1-bit GNCC or PCC, but strongly influenced by the
 237 distribution of noise sources (e.g., Snieder, 2004). Hence, to extract accurate information
 238 from amplitudes it is important to design processing flows that minimize this bias as, for
 239 example, Bowden *et al.* (2018) show in the construction of surface wave attenuation maps
 240 across the US.

241 The strongest signal we observe in Fig. 3 is the Rayleigh wave starting at the bottom-left
 242 corner and progressively vanishing at larger distance due to heterogeneities in the crust and
 243 the top upper mantle. Much weaker, we see a great variety of body waves interacting with
 244 the mantle and the core such as the P, PcP, S, ScS, PKP, PKIKP (I), or PKIKPPKIKP
 245 (I2), and of noncausal phases (Phạm *et al.*, 2018) such as PKP-PcP preceding the direct P
 246 wave arrival in the range $120 - 140^\circ$ by about 4 min. The number of station pairs per bin
 247 from 5 to 162 degrees varies from about 20 to 100 pairs; however, it reduces to a few pairs
 248 close to 180 degrees. At these interstation distances some of these signals may be interpreted
 249 as spurious arrivals from dominant sources located far from the great circle between some
 250 station pairs (Retailleau *et al.*, 2017).

251 The signals we observe in this figure are produced by a mix of ambient noise sources
 252 and coda waves from moderate and smaller earthquakes, as can be seen from the anomalous
 253 high amplitude of the ScS, PKP, and PKIKP waves. Removing anomalously correlations
 254 and sequences with high seismic activity have removed contributions from large earthquakes
 255 whose poor distribution in azimuth may introduce bias in, e.g., velocity dispersion or arrival
 256 time observations. More effective processing flows for removing coda waves can be found
 257 in Nishida (2013) and Boué *et al.* (2014). Phạm *et al.* (2018) observe some of the signals
 258 we show in Fig. 3 using coda from large earthquakes instead and explain the origin and
 259 generation mechanism of causal and noncausal phases. Further, in Fig. 3, we observe signals

260 with low apparent velocity, such as P, S, PS and PPS, mainly produced by ambient noise
261 sources.

262 Conclusions

263 We have introduced approaches that enable the application of PCC on large data volumes
264 using conventional workstations. The stand-alone code implementing PCC and 1-bit GNCC
265 is designed to be easily integrated with current processing flows, whether it is executed
266 directly or the main functions implementing each method are called from other codes.

267 We have shown that PCC simplifies to a conventional cross-correlation of unitary phasors
268 when using the power factor 2. This cross-correlation equation is fully analogous to 1-bit
269 GNCC with the single difference that 1-bit GNCC considers the sign of the real signals and
270 PCC2 the phase factor of analytic signals. Since phase carries more information than sign,
271 most applications using 1-bit GNCC should see a gain in waveform coherence and increased
272 convergence when applying PCC2.

273 By writing PCC2 as a complex cross-correlation, we have achieved a lower operation
274 complexity and, in conventional settings, about 100 times faster code than direct imple-
275 mentations. This reduces the computational cost of PCC2 to about two times that of the
276 1-bit GNCC method using only the CPU. When using other powers, GPU acceleration has
277 allowed us to apply PCC1 to the large data volume needed to extract body wave signals
278 from interstation correlations using a conventional workstation and a low-cost GPU. In gen-
279 eral, these computational cost reductions are important to efficiently find and tune the best
280 adapted methods to a given problem and, especially, to process the large data volumes
281 that modern seismic ambient noise imaging and monitoring applications demand with phase
282 cross-correlation.

283 Data and Resources

284 The data used in this study were obtained from the Incorporated Research Institutions of
285 Seismology (IRIS) and the European Integrated Data Archive (EIDA). The CPU and GPU
286 accelerated versions of code implementing 1-bit GNCC and PCC are open source under
287 the LGPL v3 license and available at <https://github.com/sergiventosa/FastPCC>. The
288 fastest Fourier transform in the west (FFTW) is available from <http://www.fftw3.org> and
289 the CUDA documentation and installation guides from [https://docs.nvidia.com/cuda/](https://docs.nvidia.com/cuda/index.html)
290 [index.html](https://docs.nvidia.com/cuda/index.html).

291 Acknowledgments

292 We thank the editors and the anonymous reviewers for improving the manuscript. This work
293 was supported by the MISTERIOS project CGL2013-48601-C2-1-R and the ANR MIMOSA
294 project ANR-14-CE01-0012. We also acknowledge support from discussions within TIDES
295 COST Actions ES1401. Numerical computations were partly performed on the S-CAPAD
296 platform, IGP, France.

297 References

- 298 Baig, A., M. Campillo, and F. Brenguier (2009). Denoising seismic noise cross correlations,
299 *J. Geophys. Res.*, **114**, no. B8, 1–12, doi:10.1029/2008JB006085.
- 300 Becker, G., and B. Knapmeyer-Endrun (2018). Crustal thickness across the Trans-European
301 Suture Zone from ambient noise autocorrelations, *Geophys. J. Int.*, **212**, no. 2, 1237–1254,
302 doi:10.1093/gji/ggx485.
- 303 Bensen, G. D., M. H. Ritzwoller, M. P. Barmin, A. L. Levshin, F. Lin, M. P. Moschetti,
304 N. M. Shapiro, and Y. Yang (2007). Processing seismic ambient noise data to obtain

305 reliable broad-band surface wave dispersion measurements, *Geophys. J. Int.*, **169**, no. 3,
306 1239–1260, doi:10.1111/j.1365-246x.2007.03374.x.

307 Boué, P., P. Poli, M. Campillo, and P. Roux (2014). Reverberations, coda waves and ambient
308 noise: Correlations at the global scale and retrieval of the deep phases, *Earth Planet. Sci.*
309 *Lett.*, **391**, 137–145, doi:10.1016/j.epsl.2014.01.047.

310 Bowden, D. C., V. C. Tsai, and F.-C. Lin (2018). Amplification and attenuation across
311 USArray using ambient noise wavefront tracking, *J. Geophys. Res.*, **122**, no. 12, 10,086–
312 10,101, doi:10.1002/2017JB014804

313 Cheng, F., J. Xia, Y. Xu, Z. Xu, and Y. Pan (2015). A new passive seismic method based
314 on seismic interferometry and multichannel analysis of surface waves, *J. Appl. Geophys.*,
315 **117**, 126–135, doi:10.1016/j.jappgeo.2015.04.005.

316 Cupillard, P., L. Stehly, and B. Romanowicz (2011). The one-bit noise correlation: a theory
317 based on the concepts of coherent and incoherent noise, *Geophys. J. Int.*, **184**, no. 3,
318 1397–1414, doi:10.1111/j.1365-246x.2010.04923.x.

319 D’Hour, V., M. Schimmel, A. F. Nascimento, J. M. Ferreira, and H. C. Lima Neto
320 (2016). Detection of subtle hydromechanical medium changes caused by a small-magnitude
321 earthquake swarm in NE brazil, *Pure Appl. Geophys.*, **173**, no. 4, 1097–1113, doi:
322 10.1007/s00024-015-1156-0.

323 Fichtner, A., L. Ermert, and A. Gokhberg (2017). Seismic noise correlation on heterogeneous
324 supercomputers, *Seismol. Res. Lett.*, **88**, no. 4, 1141–1145, doi:10.1785/0220170043.

325 Frigo, M., and S. G. Johnson (2005). The design and implementation of FFTW3, *Proc. of*
326 *the IEEE*, **93**, no. 2, 216–231, doi:10.1109/JPROC.2004.840301, special issue on “Program
327 Generation, Optimization, and Platform Adaptation”.

328 Hadziioannou, C., E. Larose, A. Baig, P. Roux, and M. Campillo (2011). Improving temporal
 329 resolution in ambient noise monitoring of seismic wave speed, *J. Geophys. Res.*, **116**, no.
 330 B7, 1–10, doi:10.1029/2011JB008200.

331 Haned, A., E. Stutzmann, M. Schimmel, S. Kiselev, A. Davaille, and A. Yelles-Chaouche
 332 (2016). Global tomography using seismic hum, *Geophys. J. Int.*, **204**, no. 2, 1222–1236,
 333 doi:10.1093/gji/ggv516.

334 Krischer, L., T. Megies, R. Barsch, M. Beyreuther, T. Lecocq, C. Caudron, and J. Wasser-
 335 mann (2015). ObsPy: a bridge for seismology into the scientific Python ecosystem, *Com-
 336 put. Sci. Disc.*, **8**, no. 1, 014,003, doi:10.1088/1749-4699/8/1/014003.

337 Lecocq, T., C. Caudron, and F. Brenguier (2014). MSNoise, a Python package for monitoring
 338 seismic velocity changes using ambient seismic noise, *Seismol. Res. Lett.*, **85**, no. 3, 715–
 339 726, doi:10.1785/0220130073.

340 Moreau, L., L. Stehly, P. Boué, Y. Lu, E. Larose, and M. Campillo (2017). Improving ambient
 341 noise correlation functions with an SVD-based Wiener filter, *Geophys. J. Int.*, **211**, no. 1,
 342 418–426, doi:10.1093/gji/ggx306.

343 Nishida, K. (2013). Global propagation of body waves revealed by cross-correlation analysis
 344 of seismic hum, *Geophys. Res. Lett.*, **40**, no. 9, 1691–1696, doi:10.1002/grl.50269.

345 Obermann, A., M. Lupi, A. Mordret, S. S. Jakobsdóttir, and S. A. Miller (2016). 3D-ambient
 346 noise Rayleigh wave tomography of Snæfellsjökull volcano, iceland, *J. Volcanol. Geoth.
 347 Res.*, **317**, 42–52, doi:10.1016/j.jvolgeores.2016.02.013.

348 Phạm, T.-S., H. Tkalčić, M. Sambridge, and B. L. N. Kennett (2018). Earth’s correlation
 349 wavefield: Late coda correlation, *Geophys. Res. Lett.*, **45**, no. 7, 3035–3042, doi:10.1002/
 350 2018GL077244.

351 Retailleau, L., P. Boué, L. Stehly, and M. Campillo (2017). Locating microseism sources
 352 using spurious arrivals in intercontinental noise correlations, *J. Geophys. Res.*, **122**, no.
 353 10, 8107–8120, doi:10.1002/2017JB014593.

354 Romero, P., and M. Schimmel (2018). Mapping the basement of the Ebro basin in Spain
 355 with seismic ambient noise autocorrelations, *J. Geophys. Res.*, **123**, no. 6, 5052–5067,
 356 doi:10.1029/2018JB015498.

357 Sánchez-Pastor, P., A. Obermann, and M. Schimmel (2018). Detecting and locating precur-
 358 sory signals during the 2011 El Hierro, Canary Islands, submarine eruption, *Geophys. Res.*
 359 *Lett.*, **45**, no. 19, 10,288–10,297, doi:10.1029/2018GL079550.

360 Schimmel, M. (1999). Phase cross-correlations: Design, comparisons, and applications, *Bull.*
 361 *Seismol. Soc. Am.*, **89**, no. 5, 1366–1378.

362 Schimmel, M., E. Stutzmann, and J. Gallart (2011). Using instantaneous phase coherence
 363 for signal extraction from ambient noise data at a local to a global scale, *Geophys. J. Int.*,
 364 **184**, no. 1, 494–506, doi:10.1111/j.1365-246X.2010.04861.x.

365 Schimmel, M., E. Stutzmann, and S. Ventosa (2018). Low-frequency ambient noise auto-
 366 correlations: waveforms and normal modes, *Seismol. Res. Lett.*, **89**, no. 4, 1488–1496,
 367 doi:10.1785/0220180027.

368 Schippkus, S., D. Zigone, G. Bokelmann, and the AlpArray Working Group (2018). Ambient-
 369 noise tomography of the wider Vienna basin region, *Geophys. J. Int.*, **215**, no. 1, 102–117,
 370 doi:10.1093/gji/ggy259.

371 Snieder, R. (2004). Extracting the Green’s function from the correlation of coda waves: A
 372 derivation based on stationary phase, *Phys. Rev. E*, **69**, 046,610, doi:10.1103/PhysRevE.
 373 69.046610.

374 Takano, T., T. Nishimura, and H. Nakahara (2017). Seismic velocity changes concentrated
 375 at the shallow structure as inferred from correlation analyses of ambient noise during
 376 volcano deformation at Izu-Oshima, Japan, *J. Geophys. Res.*, **122**, no. 8, 6721–6736, doi:
 377 10.1002/2017JB014340.

378 Taylor, G., S. Rost, and G. Houseman (2016). Crustal imaging across the North Anatolian
 379 Fault Zone from the auto-correlation of ambient seismic noise, *Geophys. Res. Lett.*, **43**,
 380 no. 6, 2502–2509, doi:10.1002/2016GL067715.

381 Tkalčić, H., and T.-S. Phạm (2018). Shear properties of Earth’s inner core constrained by
 382 a detection of J waves in global correlation wavefield, *Science*, **362**, no. 6412, 329–332,
 383 doi:10.1126/science.aau7649.

384 Ventosa, S., M. Schimmel, and E. Stutzmann (2017). Extracting surface waves, hum and
 385 normal modes: time-scale phase-weighted stack and beyond, *Geophys. J. Int.*, **211**, no. 1,
 386 30–44, doi:10.1093/gji/ggx284.

387 Young, N. (1988). *An introduction to Hilbert space*, Cambridge University Press.

388 Zeng, X., and C. Thurber (2016). A graphics processing unit implementation for time-
 389 frequency phase-weighted stacking, *Seismol. Res. Lett.*, **87**, no. 2A, 358–362, doi:10.1785/
 390 0220150192.

391 List of Figures

392	1	Computational cost of cross-correlating 649 days of data from the CAN and	
393		ECH GEOSCOPE stations using 1-bit CC, PCC2 and PCC1 with a maximum	
394		correlation lag of $\pm 12,000$ s.	19
395	2	Geographic distribution of the stations (red triangles) used in this test. . . .	20
396	3	(a) 110 million day-long interstation correlations from 39,656 station pairs	
397		sorted by distance and stacked in bins of 0.25 degrees in the period band of	
398		30 to 50 s using PCC2 and linear stacking on days with low seismic activity.	
399		(b) Day-long correlations stacked per bin.	21

Figure 1: Computational cost of cross-correlating 649 days of data from the CAN and ECH GEOSCOPE stations using 1-bit CC, PCC2 and PCC1 with a maximum correlation lag of $\pm 12,000$ s.

Figure 2: Geographic distribution of the stations (red triangles) used in this test.

Figure 3: (a) 110 million day-long interstation correlations from 39,656 station pairs sorted by distance and stacked in bins of 0.25 degrees in the period band of 30 to 50 s using PCC2 and linear stacking on days with low seismic activity. (b) Day-long correlations stacked per bin.