

Technical Notes

# Statistics of Single-Molecule Surface Enhanced Raman Scattering Signals: Fluctuation Analysis with Multiple Analyte Techniques

P. G. Etchegoin,\* M. Meyer, E. Blackie, and E. C. Le Ru†

The MacDiarmid Institute for Advanced Materials and Nanotechnology, School of Chemical and Physical Sciences, Victoria University of Wellington, P.O. Box 600 Wellington, New Zealand

The mathematical background, based on a variation of the principal component analysis (PCA) method, is developed for the understanding of fluctuating multiple analyte single-molecule (SM) surface enhanced Raman scattering (SERS) signals; with emphasis on the bianalyte SERS technique developed recently. The method and its significance are presented to provide a systematic framework with which several aspects of the statistics of SM-SERS signals can be analyzed in general. We also apply the method to a concrete example of bianalyte statistics in silver colloidal solutions and discuss related topics around experimental issues and the interpretation of single-molecule SERS data.

Single-molecule (SM) surface enhanced Raman scattering (SERS) is a subject that has attracted considerable attention in the past decade,<sup>1</sup> despite many differing views on the interpretation of the experimental data.<sup>1–3</sup> Herewith, we set to develop the necessary tools for a formal (as well as practical) analysis and understanding of single-molecule (SM) fluctuations in SERS. We shall concentrate on the statistical analysis of fluctuations in SERS active liquids as described in the recently proposed bianalyte SERS (BiASERS) technique.<sup>2</sup> This is taken as an archetypal example of the problem, but the methodology and mathematical tools can be applied to many other experimental situations; including multiple (>2) analytes and tip-enhanced Raman spectroscopy (TERS). We develop the analysis under the premise that there is *no* real absolute quantization of intensities in single-molecule SERS signals, i.e., one cannot count molecules by measuring the SERS intensity; this issue has been discussed in full detail in previous papers.<sup>3,4</sup> The bianalyte SERS technique<sup>2</sup> does provide evidence for single-molecule sensitivity and has already been exploited for

that purpose by other authors,<sup>5,6</sup> but a few questions remain; among them: (i) how much can we obtain or learn from the statistics of single-molecule events in BiASERS? and (ii) is there a systematic and reliable method for the study of single-molecule statistics in SERS from two or many analytes?

It is the purpose of this paper to develop further the ideas and tools needed to understand single-molecule fluctuations in SERS with multiple analytes. To this end, we shall develop an analysis tool based on a variation of the principal component analysis (PCA) method. We shall assume no previous exposure to the PCA method and, therefore, we shall explain the relevant steps in some detail in the Supporting Information companion material to this paper, making emphasis on the physical meaning of the results. Therein the method is tested in a “model” example of simulated data and then applied here in the main paper to real experimental examples as a demonstration. Readers who are more familiar with aspects of principal component analysis might be able to skip a large fraction of the Supporting Information to the paper. However, its presentation in full is necessary nevertheless for other authors (experienced with PCA analysis or not) to be able to reproduce and use the method proposed here.

Henceforth, the paper has two principal objectives: For a start, we aim at presenting the technique itself with its mathematical background and justification. On the other hand, we present a few experimental examples to demonstrate its use in real situations. The contents of the Supporting Information are, accordingly, somewhat independent of the specific examples discussed here in the main paper and can be used in any other experimental implementation where SM-SERS fluctuations are relevant.

## SUMMARY OF THE MODIFIED PRINCIPAL COMPONENT ANALYSIS (MPCA) METHOD

Herewith we summarize the main aspects of the MPCA method explained and developed in full in the Supporting Information to this paper. We then apply it directly to a specific experimental example and discuss its consequences. The main characteristics of MPCA are (i) it is a variation of principal

\* To whom correspondence should be addressed. E-mail: Pablo.Etchegoin@vuw.ac.nz.

† E-mail: Eric.LeRu@vuw.ac.nz.

(1) Otto, A. J. *Raman Spectrosc.* **2002**, *33*, 593.

(2) Le Ru, E. C.; Meyer, M.; Etchegoin, P. G. *J. Phys. Chem. B* **2006**, *110*, 1944.

(3) Le Ru, E. C.; Etchegoin, P. G.; Meyer, M. *J. Chem. Phys.* **2006**, *125*, 204701.

(4) Etchegoin, P. G.; Meyer, M.; Le Ru, E. C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3006.

(5) Sawai, Y.; Takimoto, B.; Nabika, H.; Ajito, K.; Murakoshi, K. *J. Am. Chem. Soc.* **2007**, *129*, 1658.

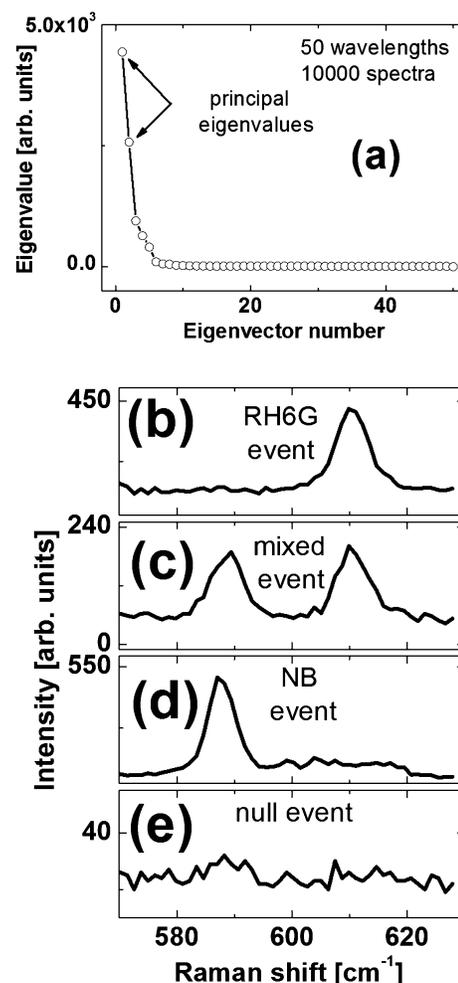
(6) Goulet, P. J. G.; Aroca, R. F. *Anal. Chem.* **2007**, *79*, 2728.

component analysis (PCA)<sup>7</sup> (widely used in many disciplines besides analytical chemistry<sup>8–10</sup>) to obtain directly the *independent components*<sup>11</sup> of the problem. The independent components are the spectra of the two different dyes in the BiASERS method and MPCA obtains them directly without going into the more sophisticated statistical analysis used in the independent component analysis (ICA) method.<sup>11</sup> The technique of ICA is, in a way, more general and powerful than PCA, for it obtains the truly independent statistical variables of the problem (the signals from two different analytes in our case). ICA is purely algorithmic also. The technique we propose here (MPCA) lays somewhere in between; it obtains the independent components of the problem but using a simpler transformation on the data. The transformation is also based on a clear physical meaning and therefore it is more transparent than the purely algorithmic versions of ICA. MPCA can be considered to be a special case of non-negative matrix factorization.<sup>12</sup> Once the independent components have been obtained, the histograms of the contributions of different dyes to the signals can be easily obtained. (ii) The acquisition of the independent components of the problem is done in MPCA through a transformation of the data provided by PCA. Accordingly, a first prerequisite is to perform a PCA analysis of the spectra. The details of the PCA analysis for this specific problem and its modification into MPCA are fully provided in the Supporting Information to this paper. We shall make explicit reference to the results in the Supporting Information for the analysis that follows in the next section.

The next section is therefore dedicated to a direct discussion of the application and interpretation of MPCA for the single-molecule SERS statistics problem in an experimental situation. We shall assume at this stage that the reader has become familiar with the mathematical concepts and ideas of MPCA developed in full in the Supporting Information.

## EXPERIMENTAL RESULTS

**Samples and Experimental Conditions.** Experiments have been performed under the same conditions as reported in ref 2. The main difference is in the type of analytes and their concentrations. The sample consists in the present case of a mixture of rhodamine 6G (RH6G) and Nile blue (NB) at 2 nM concentration each. The choice of these analytes was based on the fact that NB and RH6G have two very close Raman peaks at  $\sim 590$  and  $\sim 612$   $\text{cm}^{-1}$ , respectively. Both peaks have comparable cross sections and their proximity in energy reduces to a minimum all potential problems with the dispersion of the underlying plasmon resonance producing the enhancement.<sup>13</sup> All measurements were taken with 100 ms integration time (to minimize as much as possible temporal averaging effects due to diffusion) with a  $\times 100$  immersion objective and 3 mW laser power at 633 nm (HeNe-laser). We also used a dwell time of 1 s in between consecutive



**Figure 1.** (a) Eigenvalues (in decreasing order) obtained from PCA analysis of sample B experimental data ( $10^4$  spectra) consisting of 50 wavelengths containing the  $\sim 590$  and  $\sim 612$   $\text{cm}^{-1}$  modes of NB and RH6G. Note that the ordinate axis is not on a log scale as in Figure 4a in the Supporting Information; while the spectrum of eigenvalues is still dominated by the first two values, the next eigenvalues play a relatively more important role than in the generated model data. The spectra at the bottom show four representative examples from the  $10^4$  spectra: (b) large RH6G event, (c) medium intensity mixed event, (d) large NB event, and (e) null event (signal below the noise level). Note the different intensity scales in the different plots.

spectra (to avoid any problems with autocorrelation of intensities and to produce statistically independent events). The colloids are the conventional Lee and Meisel<sup>14</sup> citrate-reduced Ag colloids at 10 mM KCl (which have been thoroughly characterized elsewhere<sup>15</sup>) to which a premixed solution of RH6G and NB is added to a final concentration of 2 nM for each dye (sample A). We also prepared an identical sample but diluted by a factor of 3 in 10 mM KCl once it had been aggregated and the dyes incorporated (sample B).

**The Simplest Case.** With the long background of method development presented in the Supporting Information to this paper, we can proceed very quickly through the analysis of the

(7) Jolliffe, I. T. *Principal Component Analysis*; Springer: Berlin, Germany, 2002.

(8) Gemperline, P. *Practical Guide to Chemometrics*; CRC Press: New York, 2006.

(9) MacDonald, L. W.; Ronnier Luo, M. *Colour Image Science: Exploiting Digital Media*; John Wiley and Sons: New York, 2002.

(10) Bridgman, H. A.; Oliver, J. E. *The Global Climate System: Patterns, Processes, and Teleconnections*; Cambridge University Press: Cambridge, U.K., 2006.

(11) Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; John Wiley & Sons: New York, 2001.

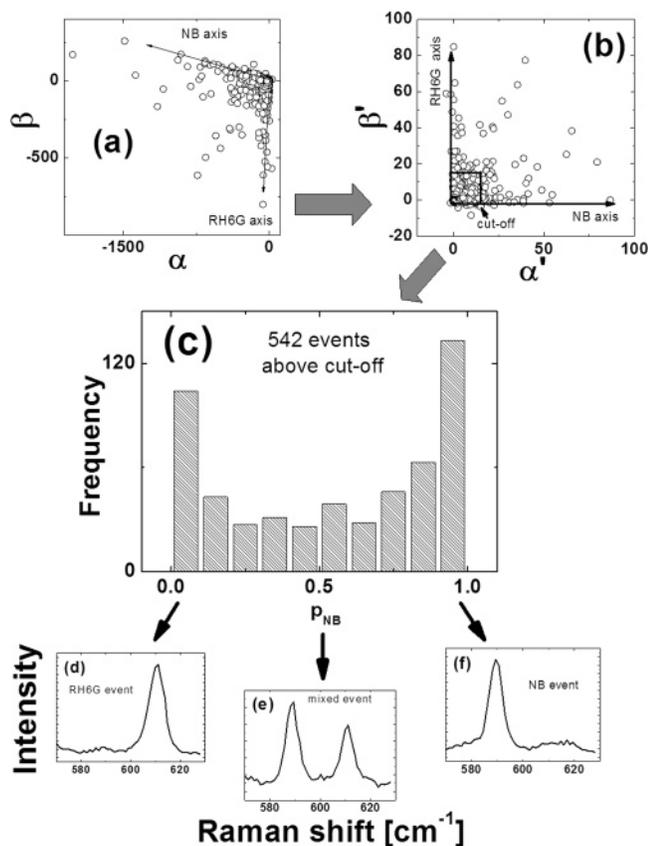
(12) Lee, D. D.; Seung, H. S. *Nature* **1999**, *401*, 788.

(13) Le Ru, E. C.; Dalley, M.; Etchegoin, P. G. *Curr. Appl. Phys.* **2006**, *6*, 411.

(14) Lee, P. C.; Meisel, D. *J. Phys. Chem.* **1982**, *86*, 3391.

(15) Meyer, M.; Le Ru, E. C.; Etchegoin, P. G. *J. Phys. Chem. B* **2006**, *110*, 6040.

(16) Le Ru, E. C.; Etchegoin, P. G. *Faraday Discuss.* **2006**, *132*, 63.



**Figure 2.** (a) Experimental **C**-matrix coefficients for sample B (with the main RH6G and NB axes indicated) with the transformed version shown in part b. The histogram in part c is obtained by discarding points below a cutoff shown in part b. Only  $\sim 5\%$  of the events can be considered to be above the noise level in this case. This histogram corresponds to a single-molecule detection regime as defined in the Supporting Information. Different spectra contributing to different regions of the histogram are shown at the bottom in parts d–f.

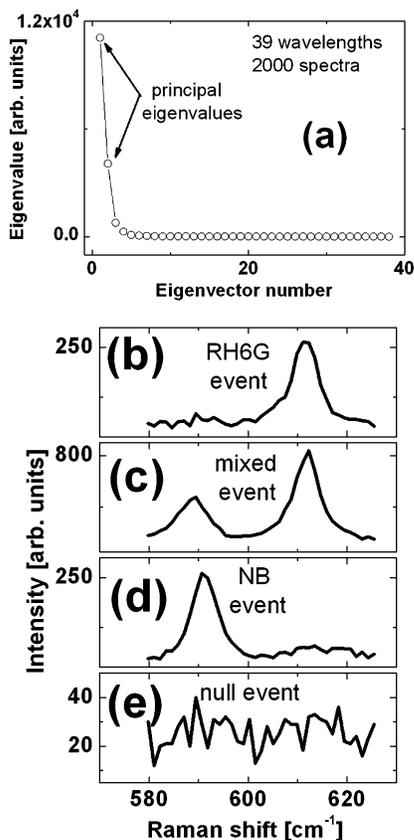
data. In this case, we analyzed first 50 wavelengths for sample B (the diluted sample) containing the  $\sim 590$  and  $\sim 612$   $\text{cm}^{-1}$  modes of NB and RH6G, respectively, and we can go directly through the intermediate steps of PCA to obtain Figure 1, where the eigenvalues of the covariance matrix are shown, together with a series of representative spectra from a total of  $10^4$ . Note that the eigenvalues spectrum is not shown on a log-scale for the ordinate (vertical) axis, as it was done in Figure 4a of the Supporting Information. This means that although the spectrum of eigenvalues is still dominated by the first two components, the influence of the next eigenvalue is comparably more important in this case than in the “model data”. We shall return to this detail in the discussion. Figure 2a shows the original **C** matrix in coefficient space (see Supporting Information), while the transformed coefficients are shown Figure 2b. The histogram in Figure 2c is obtained by counting events above a cutoff, which is decided through the same variance criterion used in the example of the Supporting Information (only  $\sim 5\%$  of the total  $10^4$  events are retained). Representative examples of spectra contributing to different regions of the histogram are shown in Figures 2d–f.

The histogram in Figure 2c has dominant contributions from the single-dye-signal events at  $p_{NB} \sim 0$  and  $\sim 1$  and therefore corresponds to the single-molecule detection regime, as discussed in Figure 6 of the Supporting Information. This is the simplest

case of SM-SERS statistics. Whenever a histogram like Figure 2c is obtained, we can be sure that the edges of the histogram (i.e., events for  $p_{NB} \sim 0$  or 1) are single-molecule events with a very high probability. We can therefore start to isolate them from the statistics and study their properties. An example of the latter will be given later in this section. Figure 2 also provides a “proof of principle” of the method developed in this paper. We know from the dye concentrations that there is approximately 20 dyes of each type per colloid;<sup>2,15</sup> for example, i.e., a maximum of  $\sim 40$  dyes of each type per active clusters (which are composed of a minimum of two colloids). The obtained histogram is therefore in agreement with the expected statistics predicted in the Supporting Information to this paper.

**Multiple Hot-Spot Averaging Effects.** Moving on from the ideal situation of the previous case with a diluted sample, we can now try to understand several experimental factors that might affect the realization of the single-molecule detection regime. The most obvious is certainly the dye concentrations. If it is higher than a threshold defined loosely by a characteristic hot-spot size, single-molecule events will be impossible to obtain. However, decreasing the dye concentration, which must always be accompanied by an increase in sampling size, may in fact be counterproductive if other factors are not taken into account carefully. The most common of these is the influence of spatial and/or temporal averaging of multiple clusters (or in a more general framework multiple hot-spots). Basically, in all the model statistics we performed in the theoretical examples there is no time limit (or experimental problem) that could prevent signals coming from a single hot-spot to be identified. In real experiments, the signal will be coming from a measurement process that may involve spatial and/or temporal averaging of different parts of the sample. In a SERS active liquid, for example, there will be a certain average number of clusters (and hot-spots) in the scattering volume (depending on colloid concentration and aggregation state) and they will be changing with a characteristic time related to the diffusion constant. Accordingly, the simplest picture of contribution from isolated single hot-spots to each “event” may have to be modified in general depending on the experiment at hand. It is precisely to avoid this problem that sample B was prepared with an additional dilution step. We can now illustrate with sample A the effect of spatial averaging when more than one hot-spot at a time is present in the scattering volume.

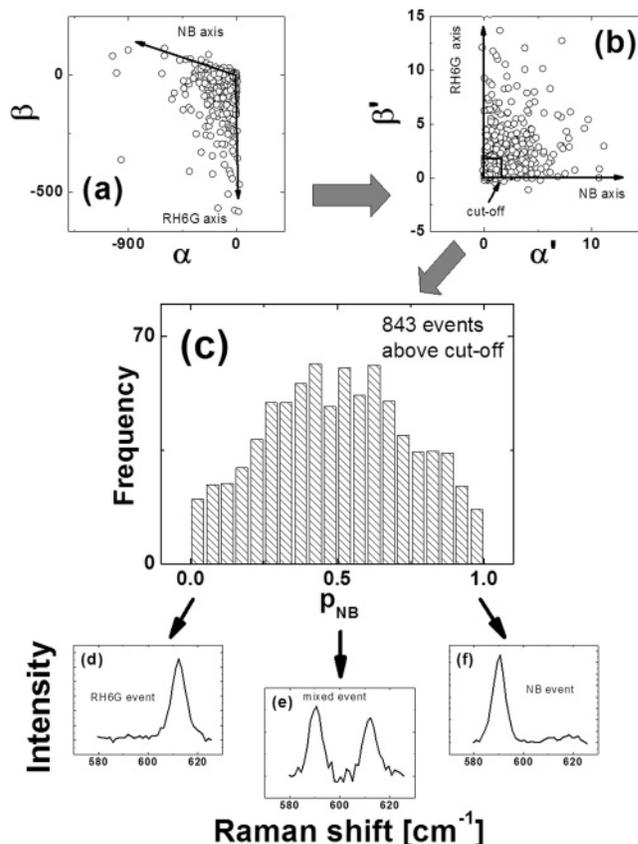
Temporal averaging has been minimized as much as possible here and in the previous experiment by choosing a small integration time compared to typical cluster diffusion times. Still, the number of hot-spots (scaling with colloid concentration) in the scattering volume needs to be taken into account (spatial averaging) even if the measurements were instantaneous. It turns out that for our experimental conditions, sample B (the diluted sample) is in the limit where single hot-spots are observed and this leads to the results of the previous subsection. Increasing the colloid concentration by a factor of 3 (sample A) yields, on the other hand, the results of Figures 3–4. Figure 3 shows the equivalent results to Figure 1, while Figure 4 shows the basic summary of the MPCA results. The histogram in Figure 4c is markedly different to that of Figure 2c, for a sample which is nominally the same but has on average three times as many clusters in the scattering volume. We are now in the few-molecule



**Figure 3.** (a) Eigenvalues (in decreasing order) obtained from PCA analysis of sample A experimental data (2000 spectra) consisting of 39 wavelengths containing the  $\sim 590$  and  $\sim 612$   $\text{cm}^{-1}$  modes of NB and RH6G. Note that, as in Figure 1a, the spectrum of eigenvalues in part a is not shown in a vertical log-scale. The spectra at the bottom show four representative examples: (b) medium-intensity RH6G event, (c) strong mixed event, (d) medium-intensity NB event, and (e) null event.

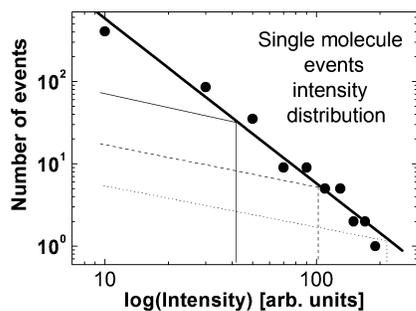
detection regime, as defined in the Supporting Information. This highlights the fact that the identification of the single-molecule detection regime is a difficult undertaking in general if only simple theoretical estimates based on dye and colloid concentrations are invoked. An unambiguous claim must be based on the detailed study of the statistics of the experimentally measured SERS events. This is where the MPCA method developed here provides a simple yet rigorous tool to carry out this type of analysis.

**Further Experimental Aspects of Single-Molecule Statistics.** Finally, there are several additional interesting aspects of the statistics of SM-SERS spectra analyzed by MPCA that are worth mentioning. (i) Real experimental data show a much more complex variety than the simulated example shown in the Supporting Information to this paper. For example, the simulated spectra results in a spectrum of eigenvalues which is dominated by two principal eigenvalues (eigenvectors), as shown in Figure 4a of the Supporting Information, with the third eigenvector being essentially noise and not accounting for any important aspect. On the other hand, the experimental data on a real BiASERS case are still dominated by two eigenvectors (eigenvalues), but to a lesser degree than the simulated example. Eigenvectors beyond the third may therefore account for small details. We find in our examples that the next four eigenvectors account for small frequency changes of the NB (3rd and 4th) and RH6G (5th and



**Figure 4.** (a) Raw coefficients of the  $\mathbf{C}$  matrix, (b) transformed coefficients, and (c) the resulting histogram coming from 843 events above cutoff, out of the 2000 events for sample A (higher cluster concentration). As before, we show in parts d–f typical spectra contributing to different regions of the histogram. The histogram indicates that we are now in a few-molecule detection regime. Many of the single-dye-signal events with  $p_{\text{NB}} \sim 0$  and 1 in this histogram are statistically “few molecules” events but not strictly single-molecule events.

6th) peaks, respectively. We note, nevertheless, that the MPCA method is not the best way to study such small frequency shifts, which can instead be obtained more accurately by direct fits and cross correlations of parameters, as done in ref 16. (ii) Many additional studies can be carried out after the present statistical analysis has been performed. Once the histograms show that we are in a single-molecule detection regime (as in Figure 2c), we can then study in more detail the single-molecule events (those close to  $p_{\text{dye1}} \sim 0$  and  $p_{\text{dye1}} \sim 1$ ). We can for example make a histogram of intensities for SM-SERS events. This provides a direct representation of the SERS enhancement factor distribution, a property which is normally very difficult to measure. This is explicitly done in Figure 5, where the long-tail nature of the average single-molecule enhancement distribution is confirmed and fitted to a Pareto distribution. The  $k$ -coefficient obtained from fits to experimental data is larger than the theoretical estimates based on a single hot-spot approach and Mie-theory calculations.<sup>3</sup> This can be understood as a manifestation of the hot-spot-to-hot-spot variability and in particular of the different maximum enhancements from one hot-spot to another. Figure 5 shows schematically how the contribution of similar distributions with different maximum enhancements (cut-offs) can result in an effective long-tail distribution with a larger  $k$  (steeper slope). What



**Figure 5.** Distribution of the SERS intensities (and therefore enhancement factors) of the single-molecule events in sample B (i.e., events with  $p_{\text{dye}1} < 0.05$  or  $p_{\text{dye}1} > 0.95$  in Figure 2c). The distribution is in a log–log plot to reveal its long-tail nature. The slope is compatible with a Pareto distribution with  $k \sim 1$ . This effective value is larger than the ones predicted for a single dimer hot-spot<sup>3</sup> because it is affected by the hot-spot-to-hot-spot variability. The thin lines illustrates how an effective distribution with a steeper slope (thick line) may arise from the multiple contributions of a collection of distributions with smaller slopes but different maximum enhancements.

the experiment provides here is the overall SM-SERS distribution with its intrinsic hot-spot-to-hot-spot variability. This approach of measuring experimentally the SERS enhancement factor distribution is quite general and can potentially be applied to a wide range of SERS substrates. (iii) Finally, one could argue that if we used ultralow concentrations, we could then simply assume that all signals are coming from single molecules and directly obtain the enhancement distribution without all the use of two analytes and the subsequent analysis presented here. While that is strictly speaking true, this approach suffers from two problems: (i) It is difficult to tell how low a concentration is required to ensure the SM-SERS nature of the signals. Only the lowest concentrations, with all the problems associated with sample preparation, would in principle be satisfactory. (ii) As the concentration is reduced, so is the probability of a molecule finding a hot-spot, which substantially reduces the number of SM-SERS events that are likely to be observed. With this in mind, the BiASERS method provides the possibility of increasing the statistical soundness of such experiments by several orders of magnitude by producing more frequent single-molecule events which can be nevertheless identified as such and counted effectively through MPCA analysis. The ideal situation to gain data from single-molecule events is therefore to increase the dye concentrations as much as possible but stop just before a situation like that shown in Figure 2c (single-

molecule detection regime) is lost. In that case, we can have the best compromise between statistical soundness (which gets better for larger number of molecules) and a situation where single-molecule events can still be identified as such. Whether spatial or temporal averaging of hot-spots is an issue is a condition that can be easily revealed in the statistics with MPCA, with a minimum effort and without any bias from specific modeling.

## CONCLUSION

This paper and the companion Supporting Information have been devoted to present the statistical analysis method of MPCA for SM-SERS spectra in the BiASERS implementation. The paper provides in addition an in-depth analysis of the statistics of single-molecule SERS signals. The method developed here should serve as an aid to the BiASERS technique to analyze data in a systematic way and, for example, to identify experimental situations in the single-molecule detection regime. This can be used in practice for the experimental measurement of the SERS enhancement factor distribution (as demonstrated in this paper). The results presented and discussed here should provide the reader with a flavor of the power of the statistical analysis based on MPCA, which is a tool not limited only to obtain the histogram of relative contributions to the total signal, and could (in addition) be easily extended to the presence of more than two analytes. We believe this to be the most comprehensive single-molecule SERS statistics study presented so far and it paves the way for future understanding and developments (of which we only presented in this paper a very basic outline of possibilities). It could provide, among other things, the basis to reveal more subtle aspects of single-molecule SERS signals involving the molecular orientation, the nature of the local field, or even interactions among different analytes.

## ACKNOWLEDGMENT

P.G.E. and E.C.L.R. acknowledge partial support for this research by the Royal Society of New Zealand (RSNZ), through a Marsden Grant.

## SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Received for review June 11, 2007. Accepted August 23, 2007.

AC071231S

# Supporting information for “Statistics of single molecule SERS signals: Fluctuation analysis with multiple analyte techniques”

P. G. Etchegoin,<sup>\*</sup> M. Meyer, E. Blackie, and E. C. Le Ru<sup>†</sup>

*The MacDiarmid Institute for Advanced Materials and Nanotechnology  
School of Chemical and Physical Sciences  
Victoria University of Wellington  
PO Box 600 Wellington, New Zealand  
(Dated: June 11, 2007)*

It is the purpose of this supporting information to the paper to present the mathematical details of the *Modified Principal Component Analysis* (MPCA) method for the analysis of single-molecule SERS fluctuations, with emphasis on the two-analyte (BiASERS) method proposed in Ref. [1] and used in the main paper here. Inevitably, the presentation of the method requires going through some preliminary information on Principal Component Analysis (PCA) itself, for MPCA is a specialization of the latter. However, rather than giving an abstract introduction to PCA we have chosen to introduce it directly with an example relevant to single molecule statistics of SERS signals, which has the main characteristic of arising from a long-tail distribution of enhancements (as studied in detail in Refs. [2, 3]). The method is presented in all its relevant details, including purely experimental issues like the removal of events that fall below a certain signal-to-noise criterion. The technique and methods studied here are somewhat independent of the content of the main paper itself, and can be used in any other relevant SM-SERS situation[4] like Langmuir-Blodgett films[5], when more than one analyte are used and the single molecule statistics of the signal becomes relevant.

PACS numbers: 78.67.-n, 78.20.Bh, 78.67.Bf, 73.20.Mf

## I. PRELIMINARIES

We first start by recalling the main ingredients of the BiASERS method with a particular attention to the statistical aspects. We also introduce a model example that can be used to describe the statistics of SERS signals at a hot-spot. This model has several advantages over previous approaches and will be used in the rest of this paper first to introduce the modified PCA method, and then to interpret the results.

### A. Definition of the problem

In a conventional single-analyte experiment, the main statistical features are related to the SERS intensity fluctuations. It is well understood now[2, 3] that the intensity fluctuations cannot be trusted as a *direct* measure of the number of molecules in SM-SERS, because of the widely diverging variety of conditions for the actual enhancement factor (EF); in other words: SM-SERS signals are *not* quantized and claims of Poisson statistics based on simple intensity analysis[6] (and obtained from a very small number of events  $\sim 100$ ) are an artifact of the limited sampling of a very skewed enhancement factor distribution[2, 3]. The latter is, in fact, a typical characteristic of high enhancement factor hot-spots[2], which

are necessary for single molecule detection.

The bi-analyte approach [1, 4, 5] makes it possible to ignore these absolute SERS intensity fluctuations, and concentrate instead on the relative SERS intensities of the two analytes. These can be analyzed, for example, in the form of a histogram of *relative contributions to the total signal* produced by a specific dye. Broadly speaking, we can consider two- (or multiple) analyte techniques as a contrast (or differential) method, in the sense that the statistics of the contribution of a single dye in a mixture is revealed in the background of the other signals. It is important however to understand that (ultimately) the BiASERS method *only* works as a proof of SM detection when single (pure) signals coming from either one analyte or the other can be pinpointed and observed. If we have a mixed signal event, we can certainly quantify the relative contribution of one analyte with respect to the other, but we cannot decide on the number of molecules producing these signals. Even signals coming from one analyte alone can suffer from the same problem at intermediate concentrations (we cannot decide if the signal is from one or a small number of molecules). It is only the combination of a contrast method like BiASERS with small analyte concentrations that ensure the single molecule nature of the signal. The statistics of SERS signals in a BiASERS experiment is therefore crucial to further confirm its interpretation in terms of single molecule detection.

We believe it is particularly important to understand how the transition from a few to single molecule events happens. From this standpoint, the main paper and this accompanying *supporting information* are about developing a better understanding of this transition in real exper-

---

<sup>\*</sup>Electronic address: Pablo.Etchegoin@vuw.ac.nz

<sup>†</sup>Electronic address: Eric.LeRu@vuw.ac.nz

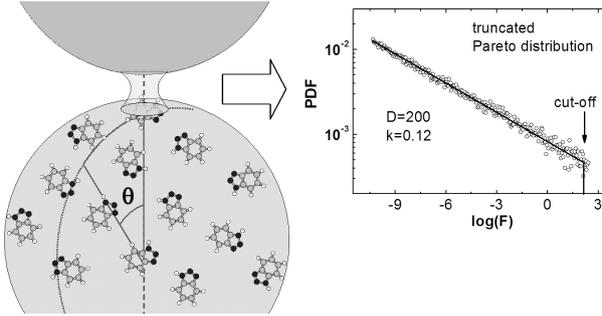


FIG. 1: For a SERS hot-spot formed between two particles (a dimer as schematically represented on the left), dyes are subject to a distribution of enhancement factors which varies drastically as a function of  $\theta$  from the maximum at  $\theta = 0$ . The probability density distribution (pdf) that results is shown on the right and was extensively studied in Ref. 2; it is a truncated Pareto distribution (long-tail). The plot shows the specific pdf we use here with  $k = 0.12$  and  $D = 200$  (See Eqs. 1-3). On top of this effect, real systems will have a hot-spot-to-hot-spot variability (typically they will have similar  $k$ 's but different cut-offs) which will result in effective parameters for the distribution. We return back to this problem in the discussion section.

imental conditions and developing the necessary mathematical tools for that purpose. We shall show that a variation of the Principal Component Analysis (PCA) method is particularly well suited for this goal, and can provide a general framework to study similar problems with two or many analytes when the SM-SERS regime is approached. In this manner, a new layer of confidence and understanding is added to the general picture of how the single molecule limit is approached in SERS, and how more general cases (including more than two analytes) can be systematically studied from a statistical point of view.

## B. A model example

Even if BiASERS substantially helps to understand SM-SERS phenomena, we still have to take into account in its interpretation the highly-skewed nature of the enhancement distribution, or in other words, the fact that SM-SERS signals originates from “hot-spots”. Hot-spots are highly localized areas on the surface, typically at the junction between two metallic objects [2, 7, 8]. One possible approach to the SM-SERS statistics at a hot-spot is to consider that a hot-spot has a given size and that

a molecule can either be inside or outside it. The statistics then arises from the probability of a molecule begin inside or outside the hot-spot. This picture is misleading since it ignores the continuous nature of the enhancement distribution and implicitly assumes “quantized” SERS intensities. We propose here a different approach that avoids these two problems. We assume that the number of molecules under consideration is fixed. The probabilistic nature of the effect is then introduced by considering that each of these molecules is subject to a random enhancement factor, which follows an appropriate enhancement distribution. This approach has a more direct link with reality, where the molecule position on the substrate is usually random, and this position determines the enhancement factor they are subject to (which can then vary continuously).

The microscopic origin of the SERS signals, therefore, is in the details of where the molecules are located on the surface with respect to hot-spots. The situation is schematically depicted in Fig. 1. The overriding effect is the *spatial* variation of the SERS enhancement factor ( $F$ ). In Ref. 2 we proposed that a good phenomenological description for the probability of the enhancement  $p(F)$  for dimers is a truncated Pareto distribution, to wit:

$$p(F) = A F^{-1-k}, \quad (1)$$

where  $A$  and  $k$  are parameters, along with the (maximum) enhancement at the hot-spot ( $\theta = 0$  in Fig. 1) which we call  $F_{\max}$ . To avoid working with large numbers, we can consider the normalized enhancement factor  $F' \equiv F/\langle F \rangle$ , where  $\langle F \rangle$  is the average (as defined in Ref. 2) given by:

$$\langle F \rangle = \frac{F_{\max}}{D}, \quad \text{with } D = \frac{1-k}{A} F_{\max}^k. \quad (2)$$

The maximum of  $F'$  is then  $D$ , as defined in Eq. 2. A random enhancement factor  $F'$ , with a truncated Pareto distribution, can be generated numerically from a variable  $U$  with uniform random distribution in the  $[0 - 1]$  range by means of:

$$F' = D \left[ 1 + \frac{k}{1-k} D(1-U) \right]^{-\frac{1}{k}}. \quad (3)$$

Note that with such an enhancement distribution, the total SERS signal of many molecules can be dominated by one molecule only, the one closest to the “hot-spot”. This is the reason why this model is particularly suited to the study of SM-SERS statistics, despite the fact that many molecules are considered.

Finally, in real cases, all hot-spots will not have the same identical parameters  $D$ ,  $k$  (and  $\langle F \rangle$ ). On top of the spatial distribution of enhancements for a single hot-spot (represented by the truncated Pareto distribution), we can expect some hot-spot-to-hot-spot variability. This will inevitably result in a quantitatively different enhancement distribution, which will nevertheless

retain its long-tail nature. If such a distribution can still be described by a Pareto distribution, as in Eq. 1, the parameters  $A$  and  $k$  must then be viewed as *effective parameters*, resulting from the hot-spot-to-hot-spot variability. A more detailed study of the variations of the parameters of the distribution with changes in incident polarization, orientation, and distance between particles has been presented in Ref. [2]. However, it is important to stress that for the present analysis the same results are obtained for as long as the final resulting distribution is long-tail (as it is in real experimental situations). *The overriding characteristic is the long-tail nature of the distribution rather than the details of its specific origin.* The choice of a Pareto distribution is based here on the fact that we showed in Ref. [2] that it has a connection with the actual electromagnetic distribution of dimers in metallic particles, but the choice is in no way a limitation to the main conclusions.

### C. Statistics for BiASERS experiments

We are now in a position to understand what the statistics of SERS signals coming from a substrate with a long-tail distribution of SERS enhancements is. The fluctuations come from the fact that, while the average number of dyes observed in each event is always the same, the specific enhancement factors for each molecule come from a long-tail random distribution and will therefore add up to different signals from one event to another. Let us assume that we have –per colloid– a certain number ( $N_1$  and  $N_2$ ) of molecules of two different types. For each molecule, we generate random enhancement factors using the model distribution defined in Eq. (3) and calculate the total intensity produced by each type of molecule ( $I_1$  and  $I_2$ ) by summing over the corresponding enhancements. From here we can derive  $I_1/(I_1 + I_2)$ , which is the fraction of the signal contributed by dye 1. If both dyes have the same intrinsic SERS cross section, the ratio  $I_1/(I_1 + I_2)$  is also *the fraction of the average number of dye 1 contributing to the signal*,  $p_{\text{dye1}}$ . We can repeat this process for a large number of events ( $T$ ), and obtain a histogram for  $p_{\text{dye1}}$ . Figure 2 shows the result for  $T = 10^5$  events with  $N_1 = N_2 = 1000, 100,$  and  $10$  molecules, respectively.

These three cases epitomize the three possible regimes of a BiASERS experiment:

- The single-molecule detection regime, as exemplified in Fig. 2(c),  $N_1 = N_2 = 10$ . In this case, the vast majority of events are of a single-dye-signal type ( $p_{\text{dye1}} = 0$  or  $p_{\text{dye1}} = 1$ ). There is still a small number of mixed-signal events  $p_{\text{dye1}} \approx 0.5$  and this is unavoidable due to the statistical nature of the effect. This indicates that a very small number of the single-dye-signal events also correspond to the signals of maybe 2 or 3 identical molecules. But the overall small number of mixed-signal tells us clearly that these are statistically negligible compared to

the real single-molecule events. The single-dye-signal events ( $p_{\text{dye1}} = 0$  or  $p_{\text{dye1}} = 1$ ) can therefore be attributed to real single-molecule events with a high probability.

- The many-molecules detection regime, as exemplified in Fig. 2(a),  $N_1 = N_2 = 1000$ . Here the absence of any single-dye-signal events ( $p_{\text{dye1}} = 0$  or  $p_{\text{dye1}} = 1$ ) clearly demonstrates that all events have contributions from many-molecules (typically more than 10).
- The few-molecules detection regime, as exemplified in Fig. 2(b),  $N_1 = N_2 = 100$ . This is the intermediate or transition regime between the previous two. The presence of many mixed-signal events suggests that many of the single-dye-signal events ( $p_{\text{dye1}} = 0$  or  $p_{\text{dye1}} = 1$ ) originate from a few ( $\sim 1-4$ ) identical molecules. It is therefore difficult to identify for sure the few true single-molecule events amongst them. However, it remains clear that the signals originate at most from a few molecules, and this regime is therefore sufficient to demonstrate unambiguously the single-molecule detection capabilities of a given SERS substrate. But for a detailed study of single-molecule events, one has to go into the single-molecule detection regime of Fig. 2(c), for example by reducing further the dye concentrations. This reduction in concentration must in practice be accompanied by an increase in the sampling (number of spectra) to retain a sound statistics.

This simple discussion highlights the importance of these types of histograms for the analysis of BiASERS experiments. The next section will therefore be devoted to developing a simple tool to directly obtain these histograms from a collection of BiASERS spectra. We will then apply this tool to real experimental data in the main accompanying paper, and will then be in a position to further discuss the interpretation of these histograms in real situations.

## II. MODIFIED PRINCIPAL COMPONENT ANALYSIS FOR SM-SERS

We now present a possible method to obtain from experimental data the statistics of  $p_{\text{dye1}}$  as in Fig. 2, and explain its main features. In order to do so, we apply it first to a set of generated data that can be controlled and changed at will. The example will provide a case where the statistics of events is sufficiently distinctive that the effect of the different potential problems (like the presence of noise) can be understood and evaluated. Last, but not least, it will introduce the Modified Principal Component Analysis (MPCA) method and several concepts that are central to the statistical interpretation of SM-SERS data in the crossover from a few to single molecule events.

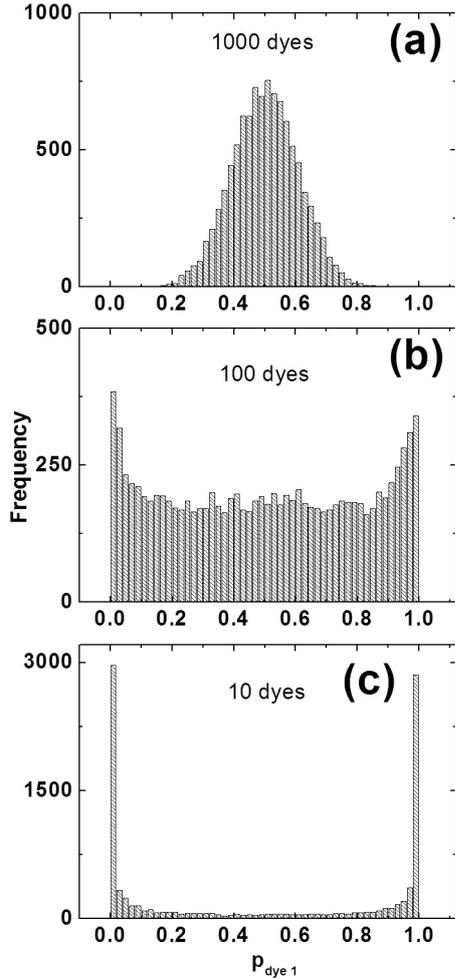


FIG. 2: Histograms of the relative contribution  $p_{\text{dye}1} = I_1/(I_1 + I_2)$  to the total intensity of dye 1 for (a)  $N_1 = N_2 = 1000$ , (b) 100, and (c) 10 molecules. Each molecule is subject to an enhancement factor with a truncated Pareto probability distribution[2]. For the normalized enhancement distribution (Eq. (3)) we used the parameters  $D = 200$  and  $k = 0.12$ , as in Fig. 1. The histogram goes from a Gaussian centered at  $p_{\text{dye}1} = 0.5$  at high concentrations in (a), to a regime where single molecule events dominate with two singularities at  $p_{\text{dye}1} = 0$  and  $p_{\text{dye}1} = 1$  in (c). The case in (b), on the other hand, exemplifies the *few molecules* transition regime in the statistics.

### A. Simulated data

We produce a set of simulated “SM-SERS model data” through the following steps, which are illustrated at the same time in Fig. 3:

- We choose a given concentration (the same) for the two dyes ( $N_1 = N_2 = 20$  in this case). We imagine

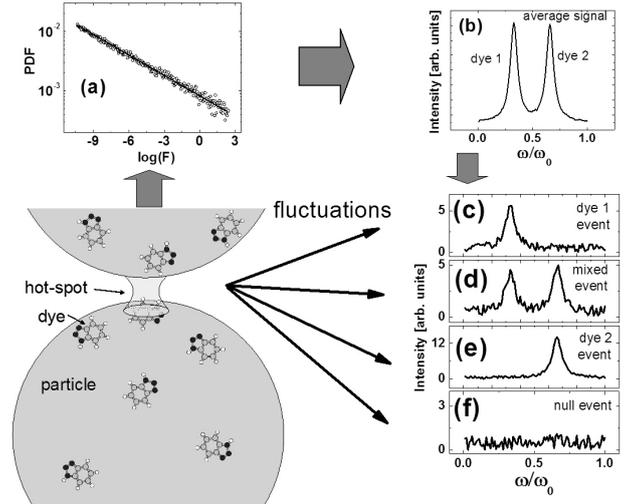


FIG. 3: Scheme for the generation of simulated BiASERS data. We choose the same number of dyes of type 1 and 2 (randomly distributed on the surface of particles, as shown on the bottom-left diagram) to be subject to the same enhancement factor distribution taken from a long-tail (Pareto) probability density (pdf) shown in (a) (with the same parameters of Fig. 1). We run the simulation for  $N_1 = N_2 = 20$ . The pdf is shown in (a) as a function of the intensity ( $I$ ) of the event in a log-log plot (as in Fig. 1). We also assume that dye 1 produces a Raman peak at  $\omega/\omega_0 = 0.33$  ( $\omega_0$  is an arbitrary scaling factor), while dye 2 has a peak at  $\omega/\omega_0 = 0.66$  and have the same SERS cross section of dye 1. Extensions to cases where the cross sections of the peaks are different are automatically taken into account in the method we develop. The average spectrum over a very large number of events (2000) in shown in (b). On the other hand, (c), (d), (e) and (f) show four representative examples of fluctuating spectra with a medium intensity dye 1 event in (c), a medium-intensity mixed event in (d), a high intensity pure dye 2 event in (e), and a null event (signal below the noise level) in (f). Random noise of a fixed amplitude is added to the spectra, to resemble real data and to demonstrate its effect on the Principal Component Analysis presented in the next subsection.

these dyes covering the surface of the particles as depicted in Fig. 3 (bottom left corner), and being subject to the same enhancement distribution, which we take as before to be a truncated Pareto distribution (Fig. 3(a)). We therefore choose a random enhancement factor for the  $N_1$  and  $N_2$  molecules from this distribution. We insist again with the remark that a Pareto distribution is taken here as an archetypal example of a long-tail distribution for the enhancement, but the results do not depend at all on this particular choice.

- We then simulate the SERS spectrum produced by these dyes. A Raman peak is generated for each dye. The SERS intensity of the peak for each dye is proportional to the sum of  $N_1$  (or  $N_2$ ) random enhancement factors. We assume that the two dyes have distinctive SERS signals (a requirement for BiASERS to work). We therefore choose arbitrarily for dye 1 to have a peak in a reduced (normalized) energy range  $0 \leq \omega/\omega_0 \leq 1$  at  $\omega_1/\omega_0 = 0.33$ , while dye two has a peak at  $\omega_2/\omega_0 = 0.66$ . We assume that the two peaks have the same intrinsic intensity (cross-section) and broadening. We then choose (arbitrarily) to have 100 wavelengths in the range  $0 \leq \omega/\omega_0 \leq 1$  and we generate a series of 2000 spectra simulating a time series of events. The average spectrum over the 2000 spectra is shown in Fig. 3(b).

- We also add noise of a fixed amplitude to the signals;  $\sim 10\%$  of the average intensity in this case. This is added explicitly to demonstrate the role and effect of noise in the analysis. Noise will be an inevitable feature in real spectra and it is important to understand how to deal with it in the statistics of events. Figures 3(c)-(f) show four representative examples of generated data. Figure 3(f), in particular, shows that signals below a certain intensity (under the noise level) will be lost.

Note that we have assumed that both dyes have the same cross sections. Different cross sections (and/or different number of molecules), however, are automatically accommodated in the general formalism we shall develop, but they are not necessary to explain the basic idea here. Moreover, the easiest experimental implementation of the BiASERS method (as far as the analysis is concerned) always comes from situations where the two dyes have distinctive peaks that are not too far away in energy (Raman shift) –to avoid unnecessary complications with SERS backgrounds[9]– and have similar cross sections. There are very many experimental implementations where these conditions are met, including the examples we shall give later in this paper.

We have now a set of spectra that simulates a BiASERS experiment for two dyes. The point at this stage is as follows: with the generated data set, we have “hidden” the original statistics of the contribution of one dye to the total intensity into a simulated set of spectra which has: (i) widely fluctuating overall intensities (coming from the long-tail enhancement distribution), and (ii) noise. The next step is to show how we invert the problem and obtain the original histogram from the simulated data. To this end, we proceed to analyze the fluctuations with the modified PCA method in the next subsections.

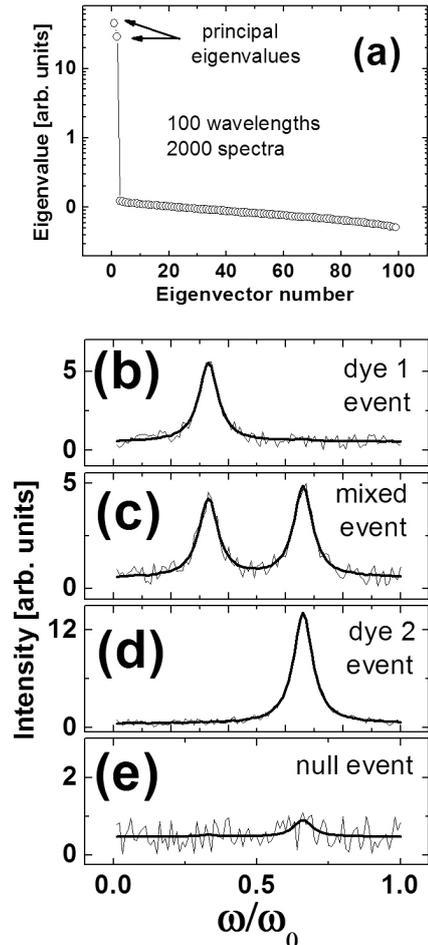


FIG. 4: (a) Eigenvalues of the covariance matrix (ordered in decreasing order). Note the logarithmic scale on the vertical axis. The first two eigenvalues completely dominate the spectrum, thus allowing a PCA-representation of the data using only two eigenvectors which we shall denote as  $f_1^{\lambda_j}$  and  $f_2^{\lambda_j}$ , respectively. In (b), (c), (d) and (e) we show the corresponding PCA-representations (thick lines) of the four events depicted before in Fig. 3(c), (d), (e) and (f), respectively (thin lines). In (f) the signal is below the noise level and the small peak in the PCA representation is an artifact of the method trying to fit fluctuations in the noise. This shows explicitly the need to exclude signals which fall below the noise level in the statistics of events.

## B. PCA analysis

The PCA method[10] (and related techniques like *Independent Component Analysis*[11]) is a well established technique with multiple applications in analytical science and spectroscopy[10]. We shall not dwell here on the technical aspects of the technique, accordingly, which are

left for the specialized literature[10, 11]. We only emphasize here the aspects of importance for the present problem.

PCA consists in applying specific linear transformations on the space defined by the “spectra”, which can then be used for “dimensionality reduction” in the data set, while retaining those characteristics of the data that contribute the most to its variance. This is done specifically by keeping only the lowest-order principal components (up to an arbitrary cut-off) and discarding the rest. Applications of PCA are wide-ranging and cover fields as dissimilar as chemometrics[12], digital image compression[13], and weather pattern recognitions[14]. PCA works well only for data that are a linear combination of independent sources, as it happens here with the additive contributions of the SERS signals from the two types of dye. For the BiASERS method, we shall be working in a situation where the first two principal components contain the essence of all the data, as we shall show later on.

We now describe the modified PCA approach to BiASERS, and illustrate its use on the data generated in the previous subsection. We define a rectangular matrix ( $M$ ) of  $T$ -times  $\times$   $N$ -wavelengths ( $T \times N = 2000 \times 100$  in this case), where we condense all the spectra as follows:

$$M = \left( \begin{array}{cccc} \overbrace{I_{t_1}^{\lambda_1} \quad I_{t_1}^{\lambda_2} \quad \dots \quad I_{t_1}^{\lambda_N}}^{N\text{-wavelengths}} \\ I_{t_2}^{\lambda_1} \quad I_{t_2}^{\lambda_2} \quad \dots \quad I_{t_2}^{\lambda_N} \\ \dots \quad \dots \quad \dots \quad \dots \\ I_{t_T}^{\lambda_1} \quad I_{t_T}^{\lambda_2} \quad \dots \quad I_{t_T}^{\lambda_N} \end{array} \right) \left. \vphantom{\begin{array}{c} \\ \\ \\ \end{array}} \right\} T\text{-times.} \quad (4)$$

Following the standard PCA implementation, we define the same matrix as before but with the mean (for each row) subtracted. This produces at each time  $t_i$  a spectrum with zero mean intensity.

$$\hat{M} = \left( \hat{I}_{t_i}^{\lambda_j} \right) \quad (5)$$

where

$$\hat{I}_{t_i}^{\lambda_j} = I_{t_i}^{\lambda_j} - \bar{I}_{t_i} \quad \text{with} \quad \bar{I}_{t_i} = \frac{1}{N} \sum_{j=1}^N I_{t_i}^{\lambda_j}. \quad (6)$$

In the next step, the covariance matrix  $V$  ( $N \times N$ ) for the  $N$  column vectors of the matrix  $\hat{M}$  ( $T \times N$ ) is calculated:

$$V = \left( \text{cov}(\hat{I}_t^{\lambda_j}, \hat{I}_t^{\lambda_k}) \right) \quad (7)$$

where  $\text{cov}(\hat{I}_t^{\lambda_j}, \hat{I}_t^{\lambda_k}) = \text{cov}(\hat{I}_t^{\lambda_k}, \hat{I}_t^{\lambda_j})$  is the covariance of the intensity columns at  $\lambda_j$  and  $\lambda_k$ , calculated here using the unbiased estimator for the covariance:

$$\text{cov}(\hat{I}_t^{\lambda_j}, \hat{I}_t^{\lambda_k}) = \sum_{i=1}^T \frac{\left( \hat{I}_{t_i}^{\lambda_j} - \langle \hat{I}^{\lambda_j} \rangle \right) \left( \hat{I}_{t_i}^{\lambda_k} - \langle \hat{I}^{\lambda_k} \rangle \right)}{(T-1)}, \quad (8)$$

with  $\langle \dots \rangle$  denoting the time-average:

$$\langle \hat{I}^{\lambda_j} \rangle = \frac{1}{T} \sum_{i=1}^T \hat{I}_{t_i}^{\lambda_j}. \quad (9)$$

Note that the covariance matrix  $V$  is a square matrix of size  $N \times N$ , i.e. its dimensionality is only defined by the number of data points in a given energy range, irrespective of the number of spectra ( $T$ ). The larger  $T$  is, however, the more accurate (in the statistical sense) the elements of the covariance matrix will be.

The next step is where the “dimensionality reduction” concept comes into play. We first obtain the  $N$  eigenvalues and  $N$  corresponding eigenvectors of the covariance matrix  $V$ . The eigenvalues are all real and positive since this matrix is real and symmetric, and can therefore be ordered from largest to smallest. The corresponding eigenvectors are the principal components in order of significance (greatest variance). The first eigenvector  $f_1^{\lambda_j}$  ( $j = 1..N$ ), for example, can be considered to be a function of wavelength that captures the most important feature in the overall set of data. In the same manner, the second eigenvector  $f_2^{\lambda_j}$  is also a function (or spectrum) that captures the second most important feature in the data, and so on. The PCA method works well (or the advantage is the greatest) in situations where we can work with a minimum number of eigenvectors.

BiASERS is a method that is particularly suited for PCA, for we shall show that we are mostly in a situation where only *two* principal components will be needed. The importance and relevance of the third eigenvector and beyond is further discussed in the last section of the main paper.

Figure 4(a) shows a plot of the eigenvalues (from largest to smallest) of the covariance matrix from the data we generated in the previous subsection. It is evident that two eigenvalues completely dominate the spectrum. This indicates that two eigenvectors will be sufficient to represent the most salient features of the data. Otherwise stated, *each individual spectrum in the time series can be to a good approximation expressed as a linear combination of the first two eigenvectors  $f_1^{\lambda_j}$  and  $f_2^{\lambda_j}$  of the covariance matrix.*

The last step is to obtain the *table of coefficients*; i.e. we need two coefficients per spectrum ( $2 \times T$  in total) in the original series that will tell us which linear combination of the first two eigenvectors  $f_1^{\lambda_j}$  and  $f_2^{\lambda_j}$  we need to represent a particular spectrum. That table is obtained from the following matrix operation (equivalent to the various scalar products of the spectra with the first two eigenvectors):

$$C = \begin{pmatrix} \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \\ \dots & \dots \\ \alpha_T & \beta_T \end{pmatrix} = \hat{M} \begin{pmatrix} f_1^{\lambda_1} & f_2^{\lambda_1} \\ f_1^{\lambda_2} & f_2^{\lambda_2} \\ \dots & \dots \\ f_1^{\lambda_N} & f_2^{\lambda_N} \end{pmatrix}. \quad (10)$$

This completes the standard PCA; it yields all the information needed to reconstruct the data: (i) the coef-

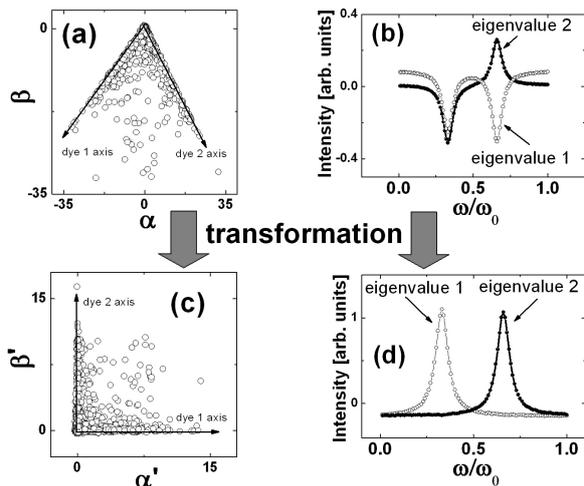


FIG. 5: In (a) we show the two dimensional (2D) representation of the matrix  $C$  in coefficient space while (b) shows the first two eigenvectors of the PCA analysis as a function of the reduced energy  $\omega/\omega_0$ . The first two eigenvectors of the covariance matrix do not necessarily separate the features we are trying to differentiate. In this particular case, for example, both eigenvectors contain a mixture of signals from dye 1 at  $\omega/\omega_0 = 0.33$  and from dye 2 at  $\omega/\omega_0 = 0.66$ . By applying a linear transformation to  $C$  through the matrix  $R$  defined in Eq. 13, we achieve the 2D representation of the coefficients shown in (c), together with the two (transformed) eigenvectors shown in (d). The principal axes for pure dye events are now perpendicular to each other and the intensities are automatically re-scaled in the transformation (which would account for possible differences in the intrinsic cross sections of the dyes and concentrations). The new (transformed) eigenvectors are now directly related to the independent contributions of the two dyes to the total signals, and a histogram of intensities can be directly obtained, as detailed in Fig. 6.

coefficients matrix  $C$  ( $T \times 2$ ), (ii) the first two eigenvectors of the covariance matrix  $f_1^{\lambda_j}$  and  $f_2^{\lambda_j}$  (each  $1 \times N$ ), and (iii) the original mean of each spectrum  $\overline{I_{t_i}}$  ( $T \times 1$ ). This represents a massive reduction on the amount of information that still captures the essential features of the data (i.e. from  $T \times N$  in the original data matrix, to  $3T + 2N$  in the final arrays).

The  $i$ -th spectrum in the time series ( $t_i$ ) is reconstructed as a function of  $\lambda_j$  as:

$$I_{t_i}^{\lambda_j} = \alpha_i f_1^{\lambda_j} + \beta_i f_2^{\lambda_j} + \overline{I_{t_i}}. \quad (11)$$

An example of “reconstruction” for the spectra shown before in Figs. 3(c)-(f) is displayed in Fig. 4(b)-(e), respectively. As can be appreciated for this latter four figures, the PCA method captures the essence and most important features of the data, ignoring the noise components and providing good quality interpolations of the data. They also show that when the intensity falls below the

noise level, the PCA representation of the signal maybe artificial. This is explicitly seen in Fig. 4(e), in which the “peak” is an artifact of the PCA representation trying to fit accidental correlations in the noise signal with only two eigenvectors. Events below certain intensity must be discarded to obtain a meaningful histogram. This leads to the concept of a cut-off in coefficient space, as we shall see in what follows.

### C. Modified PCA for BiASERS

We now specialize the PCA method for our purposes. The following points must be noted:

- For problems where two main components are needed, we always finish with two eigenvectors and a matrix of coefficients like  $C$  in Eq. 10. We can plot the two main eigenvectors  $f_1^{\lambda_j}$  and  $f_2^{\lambda_j}$  as a function of wavelength; this is done in Fig. 5. On the other hand, a matrix like  $C$  defines a *two dimensional coefficient space*. We can think of each row in  $C$  as representing a point in a plane with coordinates  $x_i \equiv \alpha_i$ , and  $y_i \equiv \beta_i$ . Each point represents an “event”. We can then plot the matrix  $C$  in coefficient space, as done also in Fig. 5.
- In Fig. 5(a) we see a clear pattern of two main axes. Since the average signal also scales with intensity, two spectra with the same relative ratio among peaks come from the same coefficients  $\alpha$  and  $\beta$  but simply re-scaled by a factor. Otherwise stated: spectra that only differ in the total intensity are obtained from a re-scaling factor of the coefficients in  $C$  and lay along lines in coefficient space. This is an important observation to count these events irrespective of their widely fluctuating intensity. *Events along lines in coefficient space are essentially the same events, only differing by their total intensity.*
- In the case at hand here, there are 2 different main axes representing pure events of dye 1 and dye 2 type, respectively. The points in between the two axes represent intermediate situations with contributions from both dyes. Points close (far away) to the origin represent weak (strong) intensity events. Points very close to the origin represent events within the noise and are mostly artificial assignments of the PCA trying to represent features in the noise with only two eigenvectors. This points are discarded for any subsequent analysis by using a cut-off to be defined later.

Still, the PCA method by itself can produce two principal eigenvectors which do *not* represent the main aspects we want to differentiate. What we want to emphasize in this case is the amount of signal we have from either dye

1 or dye 2 in the spectrum. Ideally, we want two eigenvectors that represent precisely that: the signal from dye 1 or dye 2, respectively. The PCA produces two main eigenvectors that will have (in general) a mixture of contributions of dye 1 and 2 in both eigenvectors. This is because the principal components are not based on the “physical meaning” of the signal, but rather on orthogonality and maximal variance conditions. In technical terms, this is the difference between the *principal components* and the *independent components* (PCA vs. ICA) of the problem. What we propose hereafter as Modified PCA (MPCA) method is a variation of PCA to get directly to the independent components of the problem (the individual signals of the dyes) without using the more sophisticated mathematical tools of ICA (like maximum non-Gaussianity or non-linear optimization[11]) which are more difficult to implement in general.

In terms of the physical meaning of the data, we would like: (i) the two eigenvectors to represent the actual Raman spectra, i.e. with “positive peaks” for each of the dyes, (ii) the coefficient matrix  $C$  to be composed of “positive coefficients” only, and (iii) the relative intensities of the two eigenvectors to represent exactly the dye concentrations. Although these three conditions could in principle be done by using a non-negative matrix factorization algorithm[15], we show now how this can be achieved more simply from the standard PCA analysis.

We therefore apply an appropriate linear transformation in coefficient space, to simultaneously rotate and rescale the two “pure dye” axes identified in Fig. 5(a). In more detail, we perform the following transformations on the PCA results:

- We define two vectors  $\vec{e}_1 = n_1^x \vec{e}_x + n_1^y \vec{e}_y$  and  $\vec{e}_2 = n_2^x \vec{e}_x + n_2^y \vec{e}_y$  that are two unit vectors ( $(\vec{e}_1 \cdot \vec{e}_1 = 1$  and  $\vec{e}_2 \cdot \vec{e}_2 = 1)$ ) along the principal directions representing “dye 1” events and “dye 2” events in coefficient space, as depicted schematically in Fig. 5(a). Note that the choice of  $\vec{e}_1$  and  $\vec{e}_2$  is carried out “manually” by visual inspection of the plot in Fig. 5(a). This approach is therefore not purely algorithmic, as would be a non-negative matrix factorization or independent component analysis approach, but is arguably more physical and intuitive.
- We take the average spectrum in Fig. 3(b) (with zero mean intensity) and decompose it as a sum of the two main eigenvectors of the PCA, i.e:

$$(\alpha \ \beta) = (\langle \hat{I}^{\lambda_1} \rangle \ \langle \hat{I}^{\lambda_2} \rangle \ \dots \ \langle \hat{I}^{\lambda_N} \rangle) \begin{pmatrix} f_1^{\lambda_1} & f_2^{\lambda_1} \\ f_1^{\lambda_2} & f_2^{\lambda_2} \\ \vdots & \vdots \\ f_1^{\lambda_N} & f_2^{\lambda_N} \end{pmatrix}. \quad (12)$$

We therefore obtain two coefficients  $\alpha$  and  $\beta$  (as we did for each individual spectrum in (10)), which tell us how much of the first and second eigenvector we need to represent the average.

- We now need to find the linear transformation  $R$  that rotates  $\vec{e}_1$  into  $\vec{e}_x$ , and  $\vec{e}_2$  into  $\vec{e}_y$ , with possible scaling factors. These scaling factors must be

chosen so that the transformed coefficients  $\alpha$  and  $\beta$  of the average spectrum are in the same ratio as the known dye concentrations  $c_1$  and  $c_2$  ( $c_1 = c_2$  in our example here[16]). One can show that  $R$  must then be defined as:

$$R = \begin{pmatrix} k_1 n_1^x & k_2 n_2^x \\ k_1 n_1^y & k_2 n_2^y \end{pmatrix}^{-1}, \quad (13)$$

where

$$\begin{pmatrix} k_1 \\ k_2 \end{pmatrix} = \begin{pmatrix} c_1 n_1^x & c_2 n_2^x \\ c_1 n_1^y & c_2 n_2^y \end{pmatrix}^{-1} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \quad (14)$$

- The transformation  $R$  is applied to the coefficient matrix  $C$  by standard matrix multiplication, thus defining a new table of coefficients  $C' = C ({}^t R)$ .
- The first two eigenvectors must also be transformed accordingly into  $g_1^{\lambda_j}$  and  $g_2^{\lambda_j}$  as:

$$\begin{pmatrix} g_1^{\lambda_j} \\ g_2^{\lambda_j} \end{pmatrix} = ({}^t R)^{-1} \begin{pmatrix} f_1^{\lambda_j} \\ f_2^{\lambda_j} \end{pmatrix}. \quad (15)$$

A linear transformation always maps lines into lines (and zero to zero), so we are simply reorienting and rescaling the main axes of the “fan” plot in coefficient space. This last step is what we shall call the Modified PCA (MPCA) approach. The result of applying the transformation to the example we have at hand here can be appreciated in Figs. 5(c) and (d). The advantages of performing the transformation in coefficient space is twofold:

- It enhances the two features we are trying to differentiate and count, i.e. it creates two eigenvectors which are directly linked to the presence of either one dye or the other, and they are their corresponding Raman spectra (with positive peaks), and
- it maps the “pure dye 1” and “pure dye 2” onto two perpendicular axes in coefficient space. It can also be shown that this transformation automatically rescales the intensity of one dye with respect to the other to account for cases where the intrinsic cross sections or the concentrations of the two dyes are different. The transformed coefficients are a direct measure of the *average number* of each dye (not the average SERS intensity) producing the SERS signal.

The advantages of transforming the data into this new form will be obvious now from Fig. 6. Once the transformation is performed, we can obtain directly the statistics of events independent of the total enhancement and the differences in intrinsic cross sections.  $g_1^{\lambda_j}$  and  $g_2^{\lambda_j}$  represents the average signal of a single molecule of dye 1 and dye 2 (in the same arbitrary units). We denote  $I_1$  and  $I_2$

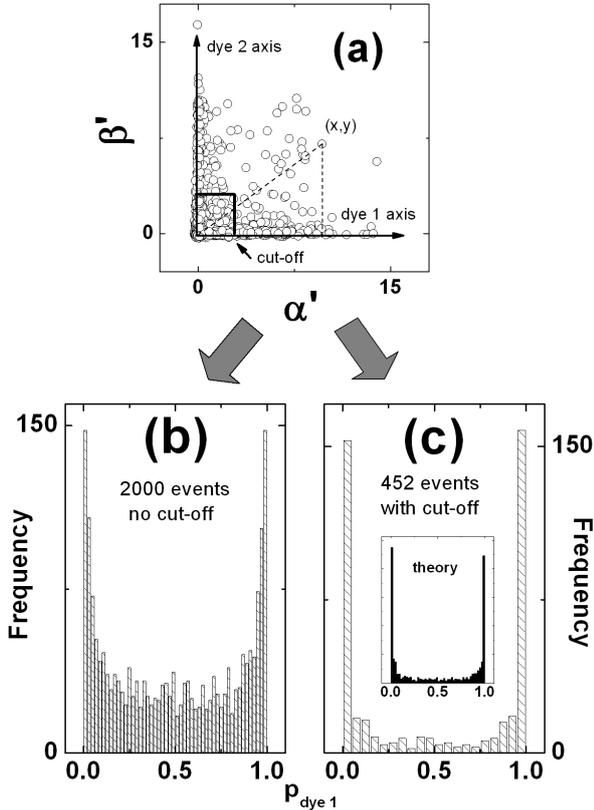


FIG. 6: Once the coefficient matrix  $C$  is transformed, the histogram of relative contributions to the total signal can be easily obtained. A point with coordinates  $(x, y)$  as shown in (a) contributes with an event of  $p_{\text{dye}_1} = 1/(1 + y/x)$ . The counting is actually done with the function  $p_{\text{dye}_1} = 1/(1 + \text{abs}(y/x))$  which avoids problems with small negative components produced by the scatter in the data along the main axes. From here, a list of events with their corresponding  $p_{\text{dye}_1}$ 's can be obtained, and a histogram can be built. The effect of noise becomes also apparent. In (b) we show the reconstructed histogram when *all* points are taken into account (including weak signals comparable to or below the noise level). The background comes from “artificial” counting of events that the PCA cannot distinguish as such below the noise level. By doing a counting of events above a certain threshold (shown in (a)) decided by a variance criterion, we recover the true histogram in (c), which agrees with the expected result for a Pareto distribution with  $N_1 = N_2 = 20$  (inset in (c)), within the expected statistical scatter. See the text for further details.

their SERS intensity. A point in coefficient space with coordinates  $x$  and  $y$  (as depicted in Fig. 6(a)) has a total intensity which is directly given by:

$$I^{\text{tot}} = (xI_1 + yI_2). \quad (16)$$

The fraction of dye 1 in terms of *signal intensity* for this event would be given by:  $xI_1/(xI_1 + yI_2)$ . However, the important quantity for analysis of a Bi-Analyte SERS experiment is the fraction of dye 1 *in terms of average*

*number of dyes* contributing to the signal for that particular event. Within our framework where different cross-sections and concentrations have already been accounted for, it is simply given now by:

$$p_{\text{dye}_1} = x/(x + y) = 1/(1 + \text{abs}(y/x)). \quad (17)$$

The use of  $\text{abs}(y/x)$  in the above expression is *a priori* irrelevant since  $x$  and  $y$  are normally positive, but it avoids problems in practice with points on the main dye axes that might go slightly negative due to the natural scatter of the statistical analysis introduced by the noise.

We can therefore obtain, for each point in the plot of Fig. 6(a), a value of  $p_{\text{dye}_1}$  and make a histogram of these values, as done in Fig. 6(b).

#### D. Removal of noisy events

Fig. 6(b) shows indeed the two singularities of the histogram we are expecting at  $p_{\text{dye}_1} = 0$  and 1, but mounted on a background. Part of this background can have a real physical origin in real experiments: due to the statistical nature of molecular adsorption, there is always a (slim) chance that one molecule of each type is located at the hot-spot, resulting in  $p_{\text{dye}_1} \approx 0.5$ . There are also issues related to the counting of multiple hot-spots, as we shall show later. However, in the model spectra at hand here, the background comes from the “counting” of many artificial cases below the noise level, where  $p_{\text{dye}_1}$  could take any random value between 0 and 1.

In order to recover the original statistics of events, we have to introduce a *threshold* in intensity at the noise level. Below this threshold, the relative intensities of the two dyes as determined by the PCA analysis, are entirely artificial and the events should not be taken into account for the statistics. If we introduce a cut-off to eliminate these cases, as shown in Fig. 6(a), we recover the original statistics of contributions of one dye to the total signal (with the inevitable scatter introduced by the PCA trimming and the finite number of spectra). This is explicitly shown in Fig. 6(c).

The cut-off can be decided with a variety of schemes, but should ultimately be validated by checking that the retained spectra are indeed above the noise. Here we chose to compare for a given spectrum the standard deviation of the error (measured signal minus its PCA representation) with respect to the intensity of the peaks (either from dye 1 or 2) measured by the PCA coefficients matrix  $C$  itself (and from the corresponding eigenvectors). Only events with a peak intensity twice above the standard deviation of the error are accepted. Other criteria are also possible and special consideration must be taken if the cross sections of the dyes are not the same. Note for example that if the dyes have different cross sections, the background is in general skewed.

### E. Summary of the MPCA approach

This completes the “proof by example”, showing that the problem can be inverted from the data using MPCA and that the statistics of contributions of one dye to the total signal can be retrieved from a time sequence of spectra. We summarize briefly what we have done in this *supplementary information*:

- We started with a clear-cut pre-defined statistics of events based on a long-tail distribution of SERS enhancements and a fixed average number of dyes for both analytes.
- We “hid” the statistics in spectra with widely varying overall intensities and background noise, as in real experiments.
- We inverted the problem and recovered the expected histogram from the statistics of signals by

using a modified PCA method (MPCA) which involves a transformation of the coefficient matrix and the two main eigenvectors.

- The statistics of events can be directly obtained from the modified coefficients and the new eigenvectors separate explicitly the two contributions we are trying to quantify.

This approach is very general and can be applied to most BiASERS data, independent of the actual nature of the SERS enhancement distribution. Its aim is to extract in a systematic and unbiased way the histogram of relative contributions to the total intensity. We have already discussed from a theoretical point of view the interpretation of these histograms in the preliminary section. In the main paper we now apply the MPCA approach to real experimental data and discuss their interpretation for a particular system: partially aggregated silver colloids.

- 
- [1] E. C. Le Ru, M. Meyer, and P. G. Etchegoin, *J. Phys. Chem. B* **110**, 1944 (2006).
- [2] E. C. Le Ru, P. G. Etchegoin, and M. Meyer, *J. Chem. Phys.* **125**, 204701 (2006).
- [3] P. G. Etchegoin, M. Meyer, and E. C. Le Ru, *Phys. Chem. Chem. Phys.* (submitted).
- [4] Y. Sawai, B. Takimoto, H. Nabika, K. Ajito, and K. Murakoshi, *J. Am. Chem. Soc.* **129**, 1658 (2007).
- [5] P. J. G. Goulet and R. F. Aroca, *Anal. Chem.* **79**, 2728 (2007).
- [6] K. Kneipp and H. Kneipp, in *Surface-Enhanced Raman Scattering: Physics and Applications (Topics in Applied Physics Vol. 103)* edited by K. Kneipp, H. Kneipp, and M. Moskovits (Springer Verlag, Berlin, 2006), p. 183.
- [7] E. C. Le Ru, C. Galloway and P. G. Etchegoin, *Phys. Chem. Chem. Phys.* **8**, 3083 (2006).
- [8] P. G. Etchegoin, C. Galloway, and E. C. Le Ru, *Phys. Chem. Chem. Phys.* **8**, 2624 (2006).
- [9] E. C. Le Ru, M. Dalley, and P. G. Etchegoin, *Curr. Appl. Phys.* **6**, 411 (2006).
- [10] I. T. Jolliffe, *Principal Component Analysis* (Springer, Berlin, 2002).
- [11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley & Sons, New York, 2001).
- [12] P. Gemperline, *Practical Guide to Chemometrics* (CRC Press, New York, 2006).
- [13] L. W. MacDonald and M. Romier Luo, *Colour Image Science: Exploiting Digital Media* (John Wiley and Sons, New York, 2002).
- [14] H. A. Bridgman and J. E. Oliver, *The Global Climate System: Patterns, Processes, and Teleconnections* (Cambridge University Press, Cambridge, 2006).
- [15] D. D. Lee and H. S. Seung, *Nature* **401**, 788 (1999).
- [16] It should be noted that implicit in this argument is the assumption that the “sticking properties” of the molecules to the metal are the same for both dyes, so that their relative importance for the statistics can be measured directly by the concentration and the intrinsic cross sections. Were this not the case, different sticking properties will appear as “effective concentrations” that take into account implicitly this difference between the surface chemical behavior of the two analytes.