Supporting information for "Statistics of single molecule SERS 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
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-

*Electronic address: Pablo.Etchegoin@vuw.ac.nz
†Electronic address: Eric.LeRu@vuw.ac.nz

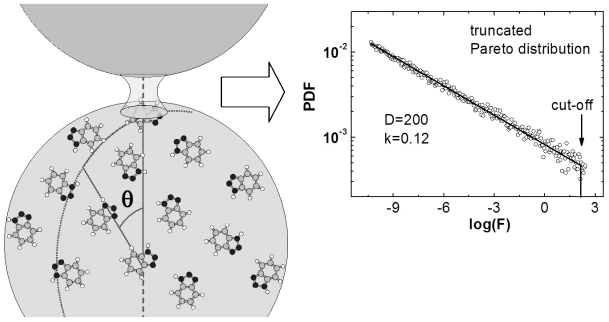

FIG. 1: For a SERS hot-spot formed between two particles (a dimer as schematically represented on the left), dyes are subject to a distribution of enhancement factors which varies drastically as a function of θ from the maximum at $\theta = 0$. The probability density distribution (pdf) that results 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:

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

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:

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

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:

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

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_{dye1}. We can repeat this process for a large number of events (T), and obtain a histogram for p_{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 ($p_{\mathrm{dye1}} = 0$ or $p_{\mathrm{dye1}} = 1$). There is still a small number of mixed-signal events $p_{\mathrm{dye1}} \approx 0.5$ 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 ($p_{\mathrm{dye1}} = 0$ or $p_{\mathrm{dye1}} = 1$) 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 ($p_{\mathrm{dye1}} = 0$ or $p_{\mathrm{dye1}} = 1$) 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 ($p_{\mathrm{dye1}} = 0$ or $p_{\mathrm{dye1}} = 1$) 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_{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.

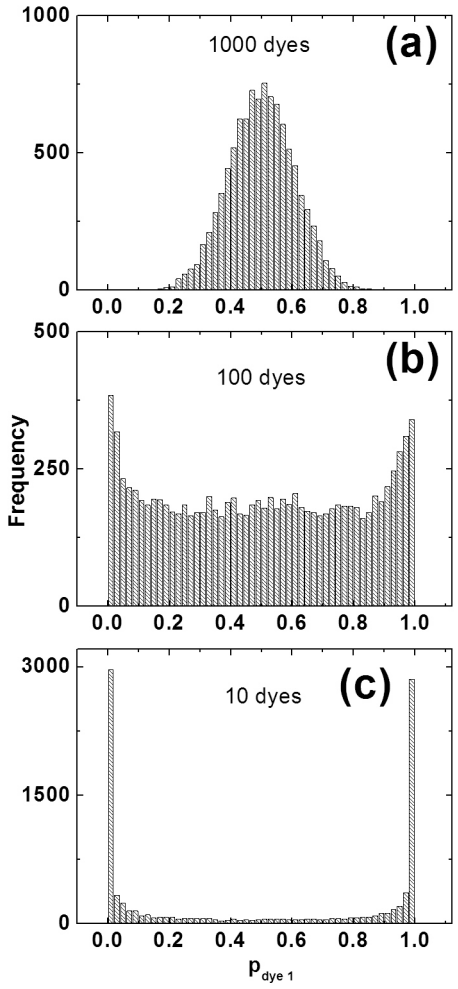

FIG. 2: Histograms of the relative contribution $p_{dye1} = I_1/(I_1 + I_2)$ to the total intensity of dye 1 for (a) $N_1 = N_2 = 1000$, (b) 100, and (c) 10 molecules. Each molecule is subject to an enhancement factor with a truncated Pareto probability distribution[2]. For the normalized enhancement 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 $p_{\rm dye1} = 0.5$ at high concentrations in (a), to a regime where single molecule events dominate with two singularities at $p_{\rm dye1} = 0$ and $p_{\rm dye1} = 1$ in (c). The case in (b), on the other hand, exemplifies the *few molecules* transition regime in the statistics.

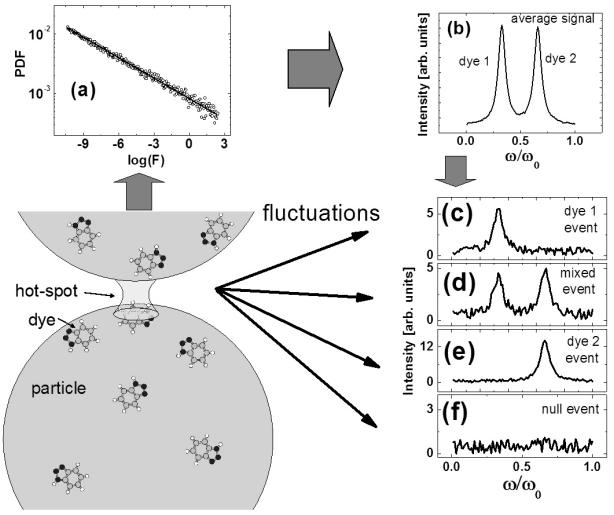

FIG. 3: Scheme for the generation of simulated BiASERS data. We choose the same number of dyes of type 1 and 2 (randomly distributed on the surface of particles, as shown on the bottom-left diagram) to be subject to the same enhancement factor distribution taken from a long-tail (Pareto) probability 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 $\omega/\omega_0 = 0.33$ (ω_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.

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 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.

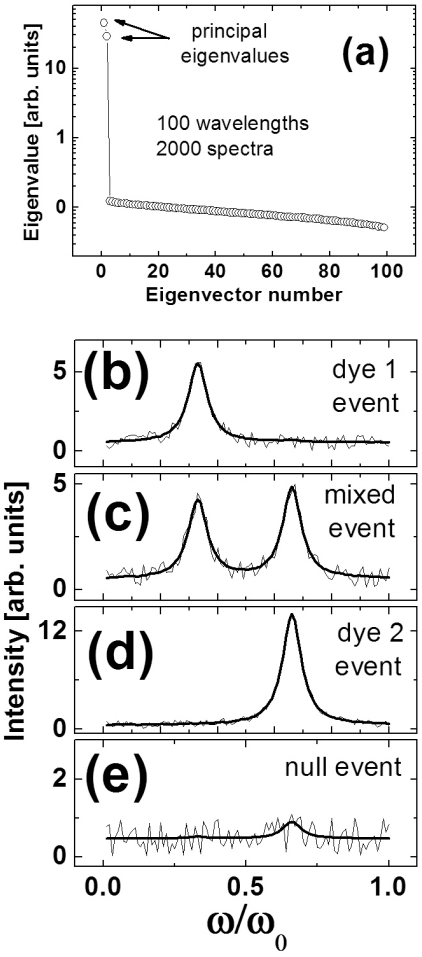

FIG. 4: (a) Eigenvalues of the covariance matrix (ordered in decreasing order). Note the logarithmic scale on the vertical axis. The first two eigenvalues completely dominate the spectrum, thus allowing a PCA-representation of the data using only two eigenvectors which we shall denote as $f_1^{\lambda_j}$ 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:

$$M = \overbrace{\begin{pmatrix} I_{t_1}^{\lambda_1} & I_{t_1}^{\lambda_2} & \dots & \dots & I_{t_1}^{\lambda_N} \\ I_{t_2}^{\lambda_1} & I_{t_2}^{\lambda_2} & \dots & \dots & I_{t_2}^{\lambda_N} \\ \dots & \dots & \dots & \dots & \dots \\ I_{t_T}^{\lambda_1} & I_{t_T}^{\lambda_2} & \dots & \dots & I_{t_T}^{\lambda_N} \end{pmatrix}}^{N-\text{wavelengths}} \Bigg\} \text{T} - \text{times}. \quad (4)$$

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

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

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 λ_j and λ_k, calculated here using the unbiased estimator for the covariance:

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

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

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

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-

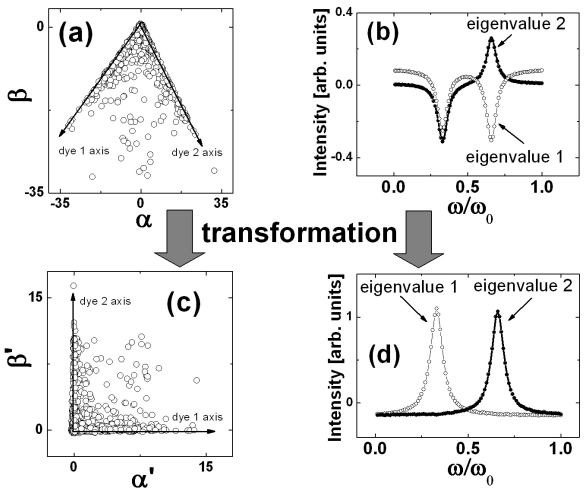

FIG. 5: In (a) we show the two dimensional (2D) representation of the matrix C in coefficient space while (b) shows the first two eigenvectors of the PCA analysis as a function of the reduced energy ω/ω_0. The first two eigenvectors of the covariance matrix do not necessarily separate the features we 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.

ficients 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 λ_j as:

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

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 α and β 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:

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

  We therefore obtain two coefficients α and β (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 α and β 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:

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

  where

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

- The transformation R is applied to the coefficient matrix C by standard matrix multiplication, thus defining a new table of coefficients $C' = C\ ({}^tR)$.

- The first two eigenvectors must also be transformed accordingly into $g_1^{\lambda_j}$ and $g_2^{\lambda_j}$ as:

$$\begin{pmatrix} g_1^{\lambda_j} \\ g_2^{\lambda_j} \end{pmatrix} = \left({}^tR\right)^{-1} \begin{pmatrix} f_1^{\lambda_j} \\ f_2^{\lambda_j} \end{pmatrix}. \qquad (15)$$

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

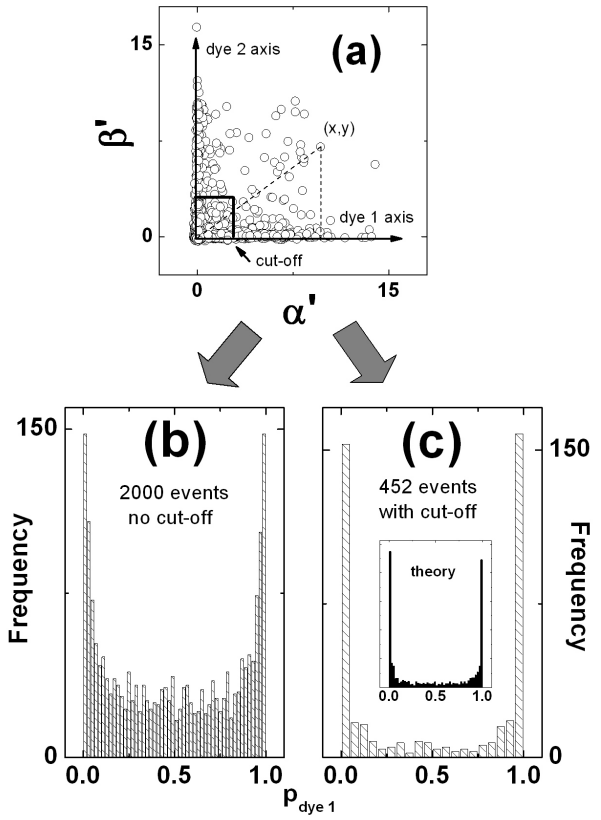

FIG. 6: Once the coefficient matrix C is transformed, the histogram of relative contributions to the total signal can be easily obtained. A point with coordinates (x, y) as shown in (a) contributes with an event of $p_{\mathrm{dye}_1} = 1/(1 + y/x)$. The counting is actually done with the function $p_{\mathrm{dye}_1} = 1/(1+\mathrm{abs}(y/x))$ 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_{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:

$$I^{\mathrm{tot}} = (xI_1 + yI_2). \tag{16}$$

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:

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

The use of $\mathrm{abs}(y/x)$ 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_{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 $p_{\mathrm{dye1}} = 0$ 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 $p_{\mathrm{dye1}} \approx 0.5$. 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_{dye1} 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. Ronnier 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.