Mixing two sets of noisy measurements changes the *N*-dependence of resolution to a fourth-root power law

2004 J. Phys. A: Math. Gen. 37 4913

(http://iopscience.iop.org/0305-4470/37/18/002)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **37** (2004) 4913–4937 PII: S0305-4470(04)75438-4

Mixing two sets of noisy measurements changes the *N*-dependence of resolution to a fourth-root power law

Alireza S Mahani, A E Carlsson and R Wessel

Physics Department, Washington University, St. Louis, MO 63130, USA

E-mail: amahani@hbar.wustl.edu

Received 31 January 2004
Published 20 April 2004
Online at stacks.iop.org/JPhysA/37/4913 (DOI: 10.1088/0305-4470/37/18/002)

Abstract

If noise is uncorrelated during repeated measurements of the same physical variable, averaging these measurements improves the accuracy of estimating the variable. When two values of a variable are measured separately, the smallest separation of these two values that can be discriminated with a certain reliability (resolution) is inversely proportional to the square root of the number of measurements employed. However, if measurements for these two values are mixed together, they need to be clustered before being averaged. Distinguishing mixed clusters with small separations can be thought of as a problem of deciding the number of components in a finite mixture model. Using the likelihood ratio, the second-moment estimator, and the k-means clustering methods, we will show that a similarly defined resolution for the mixed scenario is, approximately, inversely proportional to the fourth-root of the number of measurements. The observed fourth-root law is explained in terms of some more intuitive properties of the problem. We also conclude that, assuming that the fourth-root law is universal, the methods reported here are near-optimal.

PACS numbers: 06.20.Dk, 02.50.Cw, 02.60.Pn

1. Introduction

Selecting the number of components in fitting finite mixture models to a set of observations is an important problem that enters several types of measurements and phenomena in physics and biophysics. For example, in scattering studies of the structure of condensed matter, one is often faced with the problem of separating an angle-dependent scattering intensity into two or more peaks. Similar issues occur in magnetic-resonance studies of solids or liquids when scanning is done over magnetic field or frequency. On the other hand, the visual systems of living organisms can use biophysical machinery to decide whether they are observing one, two or more objects on the basis of space- and time-resolved signals impinging on the retina.

The optimal way of deciding the number of components has not been completely resolved (McLachlan and Peel 2000, Frayley and Raftery 1998, Hardy 1996, Tibshirani *et al* 2001, Gordon 1999, Milligan and Cooper 1985). A naive likelihood ratio procedure always favours more components, a phenomenon referred to as 'over-fitting' (Duda *et al* 2000). Two main solutions have been suggested for this problem. One solution is to subtract a penalty term from the log-likelihood that penalizes the models for their complexity, leading to what are called information criteria (Akaike 1974, Bozdogan and Sclove 1984, Sclove 1987, Schwarz 1978). (Models with more components are considered more complex.) The other main approach is hypothesis testing, using the likelihood ratio as the test statistic (McLachlan and Peel 2000, Everitt and Hand 1981). The null hypothesis is typically the model with smaller number of components. In order to provide a confidence level for accepting the alternative hypothesis, the distribution of the test statistic under the null hypothesis needs to be known.

While previous work has focused on developing, improvising and comparing different methods for selecting the number of components, in this paper we are specifically interested in how our ability to identify the correct number of components in a finite mixture model improves with the number of available measurements, regardless of which method is used for the identification process. We use simulations to study several competing methods for selecting the number of components. In order to offer quantitative statements and complement our computer simulation results with analytic theory, we choose a simplified mixture problem. We assume that measurements are independently sampled from a univariate mixture-of-Gaussians probability distribution function (PDF), with either one or two components. For the two-component case, we assume that the exact same number of measurements (N) come from either component[1]. Furthermore, we assume that the two components have equal variances (σ^2). These variances are also equal to that of the one-component Gaussian. In the one-component case, we sample $2N$ measurements, resulting in the same total as the two-component case. It is clear that if separation between the two Gaussian components is small compared to σ, it will be hard to discriminate them from a one-component Gaussian. It is also clear that as N gets larger, we will be able to reliably identify two-component PDF's with smaller separations.

To quantify the above notions, we design a *two-alternative forced choice* experiment in section 2. Based on this experiment, we can define *resolution*, which quantifies our ability to choose the correct number of components. We will introduce the *separate* scenario, only to contrast with the *mixed* scenario which is the focus of this paper. Sections 3 and 4 discuss the relationship between resolution and N for the separate and mixed scenarios, respectively. Section 5 provides a theoretical explanation for the simulation results in section 4. Finally, section 6 offers an optimality argument for the results and discusses them in the context of the related work.

2. Two-alternative forced choice experiment

We are given two sets of $2N$ measurements each. All the measurements in one set (homogeneous set) are sampled from a one-component Gaussian (σ^2), and all the measurements in the other set (inhomogeneous set) are sampled, in equal numbers, from two Gaussians (σ^2) with a mean separation of δ. (We can say that, for the homogeneous set, $\delta = 0$.) After sampling the measurements for the inhomogeneous set, they are mixed

[1] Technically speaking, sampling $2N$ measurements from $\frac{1}{2}[f_1(x) + f_2(x)]$ is not the same as sampling N measurements from $f_1(x)$ and $f_2(x)$ each. The former does not result in the exactly same number of measurements coming from each individual component. The latter is our assumption while the former is more typical of a mixture problem. The distinction, however, is only academic in this paper as these two settings provide almost identical results.

together. The two sets are unlabelled and our task is to label them, i.e. to determine which set is homogeneous and which set is inhomogeneous, using any of the methods described later in this paper. This is called a two-alternative forced choice because we have two labels to assign to two sets, and there are only two different ways to do this. In other words, for a given δ and N, we may or may not label the two sets correctly. We thus define the probability P as the fraction of correct labelling instances if the experiment is repeated an infinite number of times, with random measurements generated independently from the underlying PDF's each time.

In our simulations, we use the results of a finite but large number (10 000) of runs to approximate P. In each run, we generate $4N$ random numbers from Gaussian distributions of variance σ^2 as follows: three subsets consisting of N numbers and each is generated from the same Gaussian, while the fourth subset is generated from a Gaussian whose mean is shifted by δ with respect to the first Gaussian. The first two subsets comprise the homogeneous set, while the remaining two subsets (which are sampled from Gaussians of different means) make up the inhomogeneous set. We use one of the three methods described in detail in section 4 to label the sets.

Clearly, P is a monotonically increasing function of δ and N. If $\delta = 0$ (both sets are sampled from a single Gaussian), we expect $P = 1/2$ for any N. If $\delta \gg \sigma$, on the other hand, we expect $P \to 1$. In this paper, we are interested in how δ changes with N, for a fixed value of P.

For the sake of comparison, we also conceive of an alternative scenario which is very similar to the above, except that in the inhomogeneous set, the measurements sampled from the two Gaussians are **not** mixed. In other words, individual measurements are labelled, forming two subsets containing N measurements each. The homogeneous set is also presented in the form of two N-measurement subsets. Again, we have to decide which set is homogeneous and which set is not. We will refer to this arrangement as the separate scenario.

The advantage of using the two-alternative forced choice paradigm is that it combines the 'type I' and 'type II' errors, related to the false alarm and hit rates, into a single measure of performance (P) (Dayan and Abbott 2001). However, it should be noted that P, as defined here, is different from its typical definition in hypothesis testing, which is only related to the hit rate.

For the separate case, we expect that $\delta \propto N^{-1/2}$ (see the following section), but what about the mixed case? A faster rate of improvement is unlikely, but a slower rate is definitely a possibility. Evaluating the significance of the difference between the mixed versus separate scenarios will be the subject of this paper.

3. Separate sets

First we need to have a decision rule for labelling the two sets (homogeneous and inhomogeneous). The obvious solution is to calculate the mean of the measurements in each of the four subsets and subtract the two means for each set. The set with the larger absolute value of the difference is labelled inhomogeneous, and the other set is labelled homogeneous.

Consider one set and assume that its two Gaussians (from each which N measurements are sampled) have means of μ_1 and μ_2. We also define $\delta \triangleq |\mu_2 - \mu_1|$. (For the homogeneous set, $\delta = 0$.) Similarly, denote the measured subset means by $\hat{\mu}_1$ and $\hat{\mu}_2$, and define $\hat{\delta} = |\hat{\mu}_2 - \hat{\mu}_1|$. The pdf describing $\hat{\delta}$ is given by

$$p(\hat{\delta}|\delta) = \frac{1}{\sigma'\sqrt{2\pi}} \left\{ \exp\left[-\frac{(\hat{\delta} - \delta)^2}{2\sigma'^2} \right] + \exp\left[-\frac{(\hat{\delta} + \delta)^2}{2\sigma'^2} \right] \right\} \qquad \text{if} \quad \hat{\delta} \geqslant 0$$
$$= 0 \qquad \text{if} \quad \hat{\delta} < 0 \tag{1}$$

with $\sigma' = \sigma\sqrt{\frac{2}{N}}$.

The probability (P) of an inhomogeneous set with absolute mean difference δ being correctly discriminated from a homogeneous set (with $\delta = 0$) is equal to the probability of a random sample drawn from $p(\hat{\delta}|\delta)$ being larger than a random sample drawn from $p(\hat{\delta}|0)$:

$$P(\delta) = \int_{-\infty}^{+\infty} \mathrm{d}u'\, p(u'|0) \int_{u'}^{+\infty} \mathrm{d}v'\, p(v'|\delta) \tag{2}$$

$$= \frac{1}{\pi\sigma'^2} \int_0^{+\infty} \mathrm{d}u' \exp\left(-\frac{u'^2}{2\sigma'^2}\right) \int_{u'}^{+\infty} \mathrm{d}v' \left\{ \exp\left[-\frac{(v'-\delta)^2}{2\sigma'^2}\right] + \exp\left[-\frac{(v'+\sigma')^2}{2\sigma'^2}\right]\right\}$$

$$= \frac{1}{\pi} \int_0^{+\infty} \mathrm{d}u' \exp(-u'^2/2) \int_{u'}^{+\infty} \mathrm{d}v' \left\{ \exp\left[-\left(v'-\frac{\delta}{\sigma'}\right)^2 \Big/ 2\right] + \exp\left[-\left(v'+\frac{\delta}{\sigma'}\right)^2 \Big/ 2\right]\right\}. \tag{3}$$

The above expression is only a function of $\frac{\delta}{\sigma'}$. (The specific form of this function is not important for our discussion.) Since $\sigma' = \sqrt{\frac{2}{N}}\sigma$, P is a function of $\frac{\delta\sqrt{N}}{\sigma}$. Therefore, for a given P, $\frac{\delta}{\sigma} \propto N^{-1/2}$.

4. Mixed sets

The previous definitions for μ_1, μ_2 and δ are valid here as well. Again, we need to come up with a decision rule for how to label the two sets as homogeneous or inhomogeneous. We will discuss several such decision rules and the N–δ relationship resulting from each.

4.1. Likelihood ratio method

We calculate the maximum likelihoods (MLs) for each set assuming one and two components, and form the ratio of one-component to two-component likelihoods for them, or equivalently, subtract the log-likelihood terms. We will then choose the set with the larger ratio to be the inhomogeneous one. This is because the set with the larger two- to one-component ratio shows a larger increase in goodness-of-fit upon transition from the one-component model to the two-component model and is therefore more likely to have actually come from a two-component pdf.

In our simple problem, the (log) likelihood ratio will have the following form:

$$T = \frac{1}{N}\left\{\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{2\sigma^2} + \sup_{\hat{\mu}_1,\hat{\mu}_2} \sum_{i=1}^{N} \log\left[\exp\left(-\frac{(x_i-\hat{\mu}_1)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x_i-\hat{\mu}_2)^2}{2\sigma^2}\right)\right]\right\}. \tag{4}$$

where $\bar{x} = (x_1 + x_2 + \cdots + x_N)/N$. The $\frac{1}{N}$ factor is included to ensure that the mean of T independent of N. We have otherwise ignored the constant additive or multiplicative terms since they do not affect the outcome of a two-alternative forced choice.

In practice, finding $\hat{\mu}_1$ and $\hat{\mu}_2$ that maximize the second term in the above expression for T is not easy. The expectation-maximization (EM) algorithm (Dempster *et al* 1977, McLachlan and Krishnan 1997) finds local maxima in the likelihood function. As such, it is not guaranteed to find the global maximum, and in fact, when component means are too close compared to their spread, this algorithm performs poorly (Everitt and Hand 1981,

Titterington *et al* 1985). Since in our problem the focus will on close components, EM is not the best option.

Another iterative algorithm, which can be considered an approximation to the EM algorithm, is *k-means clustering* (Lloyd 1982). In our problem, k-means clustering can be described as follows:

(i) Make initial estimates of μ_1 and μ_2. We will refer to these estimates as $\hat{\mu}_1$ and $\hat{\mu}_2$.
(ii) Assign each measurement to the estimate to which it is closer. This splits the set of measurements into two clusters, C_1 and C_2: $C_1 = \{x_i | |x_i - \hat{\mu}_1| < |x_i - \hat{\mu}_2|\}, C_2 = \{x_i | |x_i - \hat{\mu}_1| \geqslant |x_i - \hat{\mu}_2|\}$.
(iii) Renew estimates by averaging the measurements in each cluster: $\hat{\mu}_1 = \langle x_i \rangle_{C_1}$, $\hat{\mu}_2 = \langle x_i \rangle_{C_2}$.
(iv) If converged, stop. Otherwise, go back to 2.

The main difference between EM and k-means clustering is that in the latter, probabilities for each measurement belonging to either components are replaced by 0 or 1, depending on which current estimated component mean the measurement is closer to.

We observe that if we first apply the k-means clustering to the data, and then use the results to initialize EM, the likelihood function does not improve significantly, and it can actually decrease. Moreover, k-means clustering does not suffer from the slow convergence of the EM method for close components. For these reasons, we take the output of k-means clustering as the approximate ML-estimate.

We can now use different component-mean separations (δ) and find the corresponding P-values for them, following the procedure described in section 2. We adjust δ until P is close enough to the fixed value that we desire (0.9, for example). The δ we find is valid for a certain N. We repeat this process for different values of N, resulting in a N–δ data set, which is shown in figure 1. If we fit a curve of the form $\delta = AN^{\beta}$ to these data, we find that $A = (3.27 \pm 0.25)\sigma$, $\beta = -0.271 \pm 0.015$. In other words, roughly speaking, $\delta \propto N^{-1/4}$. If we choose a different value of P, A changes but β does not.

4.2. Information theoretic methods

Information theoretic approaches such as Akaike's information criterion (AIC) (Akaike 1974, Bozdogan and Sclove 1984, Sclove 1987) or the Bayesian information criterion (BIC) (Schwarz 1978, Frayley and Raftery 1998) add a penalty term to the log-likelihood, that is a function of d, the number of parameters in the model and/or N, the number of measurements. Neither of these two parameters depend on the measurement set. Therefore, the likelihood ratio will only take up a constant term which has no effect on the two-alternative forced choice we are considering. In other words, as far as our problem is concerned, information theoretic methods will perform exactly the same as likelihood ratio approaches.

4.3. Using estimators of δ as criteria in the two-alternative forced choice

There is no reason to limit ourselves to the likelihood ratio as the statistic used in the two-alternative forced choice. Any other statistic whose mean is a monotonic function of δ can be used instead. An obvious choice is the use of estimators that are designed to reproduce δ. Here we consider two of them.

4.3.1. Second-moment estimator method. The *method of moments* (Pearson 1894, McLachlan and Peel 2000) approximates the (theoretical) mixture pdf moments with the

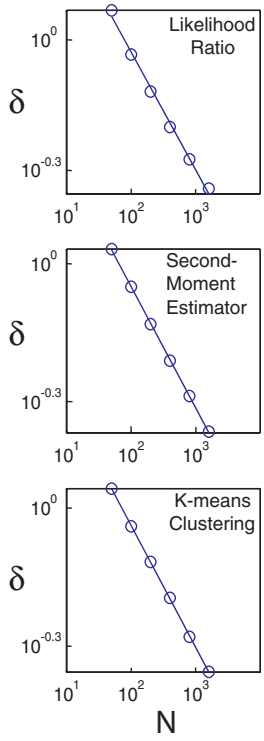

Figure 1. Log–log plots of the separation between the component Gaussian means (δ) as a function of the subset size N, estimated through simulations, for the three methods used in the mixed case (circles) and the regressed lines; the success rate P is 0.9 for all data points. δ is given in units of standard deviation σ of component Gaussians.

(empirical) moments of the measurements to find the parameters of the model. In our simple problem, the second moment of the measurements (V_2) suffices to estimate δ:

$$\hat{\delta} = 2(V_2 - \sigma^2)^{1/2}$$

with

$$V_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{5}$$

where x_i are the measurements and $\hat{\delta}$ is our estimate of δ. A problem with the above solution is that, for small cluster separations, V_2 can be smaller than σ^2, resulting in imaginary $\hat{\delta}$. To avoid this, we can compare $\hat{\delta}^2$ instead of $\hat{\delta}$, allowing for negative values as well. In fact, since any monotonic transformation of a statistic does not affect the outcome of the two-alternative forced choice, we can use $V_2^{1/2}$ as our statistic. (The set with a larger $V_2^{1/2}$ will be labelled inhomogeneous.) Besides avoiding imaginary results, this transformation also brings the pdf of our statistic closer to a Gaussian shape. The significance of this will be clarified later in the paper. Finally, evaluating $V_2^{1/2}$ does not require prior knowledge of σ.

As before, we find δ for different N and regress a straight line in the log–log plot of N–δ, as shown in figure 1. The results are $A = (3.00 \pm 0.05)\sigma$ and $\beta = -0.264 \pm 0.003$, which are very close to those from the likelihood ratio method.

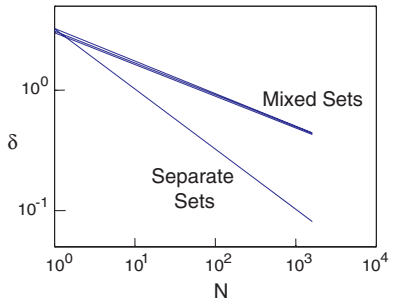

Figure 2. Log–log plots of N–δ for the separate scenario, and for three different methods used in the mixed scenario; $P = 0.9$; δ is given in units of σ; 10 000 runs per N are used to estimate δ.

4.3.2. k-means clustering method. Instead of using the result of this method to form the likelihood ratio, as we did earlier in this section, we can simply use its estimate of δ, i.e. $\hat{\delta} = |\hat{\mu}_2 - \hat{\mu}_1|$, as the statistic and, as in the second-moment estimator method, choose the set with the larger $\hat{\delta}$ as inhomogeneous. The regression results for the N–δ data in this case are $A = (3.10 \pm 0.04)\sigma$ and $\beta = -0.265 \pm 0.002$ (figure 1).

4.4. Section summary

Two points in the above results merit attention. First, the N–δ data from different methods are very close, and for all these methods, we have, approximately, $\delta \propto N^{-1/4}$. We will refer to this relationship as the *fourth-root law*. Figure 2 summarizes the N–δ data for the separate-set as well as the mixed-set scenario. Note that the regression used data only for $50 \leqslant N \leqslant 1600$ (6 points equally spaced in log space). We show the fit over a larger range in order to better contrast the slopes of the lines, and to highlight the approximate intersection points of the mixed-set and separate-set lines. We will talk about the significance of this latter point in the discussion.

The second point is that, even for a modest $N = 50$, the δ found for $P = 0.9$ is about $1.1\,\sigma$, which is about a half of the minimum separation for the mixture pdf to have two maxima. This is also a small enough separation for the EM algorithm to perform poorly, confirming our assertion that the EM method is not suitable for this problem.

5. Explaining the fourth-root law

All the methods we used in the previous section follow the same pattern: they compare a function of the measurements (the test statistic) in the two sets and choose the set with the larger number to be the inhomogeneous set. The particular function is different in each case, but they share some properties that we will discuss in this section. We refer to this function as R. Since this function operates on a random set of numbers, it is itself a random number. Moreover, its pdf is invariant under an equal shift in component means. Therefore, we can completely parametrize this pdf by N and δ, and we denote it by $p(R|\delta; N)$.

In this section, we will prove that the following premises will result in the fourth-root law. Note that this is a sufficient but not necessary set of conditions.

(i) The pdf $p(R|\delta; N)$ is Gaussian.
(ii) The variance $\left(\sigma_R^2\right)$ of R approaches a finite non-zero limit as $\delta \to 0$, for a fixed N.
(iii) $\sigma_R^2 = \mathrm{Var}[R] \propto \frac{1}{N}$, for a fixed δ.

(iv) For small δ, the mean m of R is a quadratic function of δ: $m(\delta) = \mathrm{E}[R] = a + c\delta^2$.

The first premise means that

$$p(R|\delta) = \frac{1}{\sqrt{2\pi}\sigma_R(\delta)} \exp\left(-\frac{(R - m(\delta))^2}{2\sigma_R^2(\delta)}\right) \tag{6}$$

where the dependence on N has been suppressed for now. Putting the above expression for $p(R|\delta)$ in equation (2) (R plays the role of $\hat{\delta}$), we find

$$P(\delta, N) = \frac{1}{2\pi\sigma_0\sigma_R(\delta)} \int_{-\infty}^{+\infty} \mathrm{d}u' \exp\left(-\frac{(u' - m_0)^2}{2\sigma_0^2}\right) \int_{u'}^{+\infty} \mathrm{d}v' \exp\left(-\frac{(v' - m(\delta))^2}{2\sigma_R^2(\delta)}\right) \tag{7}$$

with $m_0 = m(0)$ and $\sigma_0 = \sigma_R(0)$. We have suppressed the N-dependence of the terms on the right-hand side. Two simple changes of variables give

$$P(\delta) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{d}v \exp\left(-\frac{v^2}{2}\right) \int_{\frac{\sigma_0}{\sigma_R(\delta)}v + \frac{m_0 - m(\delta)}{\sigma_R(\delta)}}^{+\infty} \mathrm{d}u \exp\left(-\frac{u^2}{2}\right). \tag{8}$$

The above form makes it clear that $P(\delta, N)$ is only a function of $\frac{\sigma_0}{\sigma_R(\delta)}$ and $\frac{m_0 - m(\delta)}{\sigma_R(\delta)}$. To make their dependence on N explicit, we say that $P(\delta, N)$ is a function of $\frac{\sigma_0(N)}{\sigma_R(\delta,N)}$ and $\frac{m_0 - m_R(\delta)}{\sigma_R(\delta,N)}$. Premises 2 and 3 result in the first fraction, $\frac{\hat{\sigma}_0(N)}{\hat{\sigma}(\delta,N)}$, approaching unity as $N \to \infty$. As for the second fraction, $\frac{\hat{m}_0 - \hat{m}(\delta)}{\hat{\sigma}(\delta,N)}$, the denominator is proportional to $N^{-1/2}$. To make this fraction (and thus P) independent of N, we then need $m_0 - m(\delta) \propto N^{-1/2}$. But from the fourth premise, $m_0 - m(\delta) = c\delta^2$. Therefore, we need $\delta^2 \propto N^{-1/2}$, which means $\delta \propto N^{-1/4}$.

The rest of this section will discuss the validity of each of the above four premises. More detailed proofs are given in appendices B and C.

5.1. Gaussianity of $p(R|\delta)$

Figure 3 shows normalized histograms for $R(\delta = 0.5, N = 50)$ for all the three methods, each accompanied by a regressed Gaussian. It is clear that while Gaussianity is a very good approximation for the second-moment estimator and k-means clustering methods, it is invalid for the likelihood ratio method. It is possible to modify the likelihood ratio statistic to bring its distribution closer to Gaussianity, but remember that there is nothing fundamentally wrong with using a test statistic with a non-Gaussian pdf, because we are interested not in calculating a confidence level for a given set of measurements, but rather assessing the dependence of the resolution on N. We have already seen that the resulting N–δ data for the likelihood ratio method obeys the same power law as the other methods, but of course proving it would be much more difficult given the non-Gaussian distribution of R in this case.

The deviation from Gaussianity for the likelihood ratio method, as illustrated in figure 3, is not extreme. Moreover, for larger N, such deviations become less significant. Therefore, we will continue to include simulation results for the likelihood ratio method alongside other methods in our discussion of other premises, but keeping in mind that the whole argument for the likelihood ratio will have an added approximation due to the non-Gaussianity of $p(R|\delta; N)$. The proofs in appendices C and B, however, will exclude the likelihood ratio method.

5.2. Dependence of σ_R on δ

Consider the resolution values for $N = 100$ and $N = 200$, with $K = 0.9$ in both cases. For each N and its corresponding resolution, we find the estimate error, σ_R. As figure 4 illustrates for all three methods, σ_R approaches a finite N-dependent value as $\delta \to 0$.

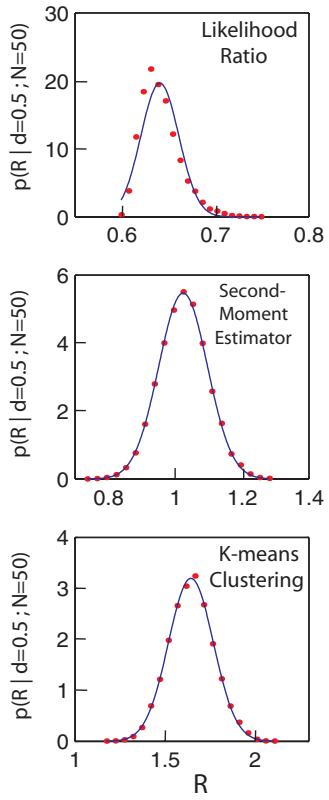

Figure 3. Probability distribution of R for $\delta = 0.5$ (normalized by σ) and $N = 50$, for all three methods; circles: empirical values found through simulation; curves: fitted Gaussians.

5.3. Dependence of σ_R on N

The third premise is justified by numerical results as well as proofs for special cases. Figure 5 shows the log–log plots of $\sigma_R(\delta = 0.5)$ as a function of N, using all three methods. A linear regression shows that the exponents are very close to -0.5. In appendix B.1 we prove this premise for the special case of a Gaussian mixture pdf using *naive estimation*, a single-pass approximation to the k-means clustering. (This essentially restricts the proof of this premise to $\delta = 0$.) The naive estimation method is explained in appendix A. Even for this very simplified situation, the proof is lengthy. A proof for the second-moment estimator method is also presented in appendix B.2.

5.4. Dependence of m on δ

If m is an even function of δ, i.e. if $m(\delta) = m(-\delta)$, and if $m(\delta)$ is a twice differentiable function, the parabolic shape will follow. Although proving this second condition is not easy when an iterative method such as the k-means clustering is used, the above argument suggests why the fourth-root law might be general. In appendix C, individual proofs for the second-moment estimator and naive estimation methods are presented. Figure 6 shows the simulation results for $N = 200$. The quadratic behaviour of m is evident.

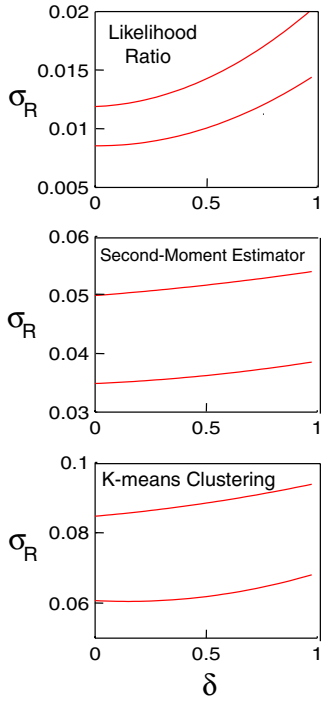

Figure 4. Smoothed plots of δ–σ_R for $N = 100$ (upper curves) and $N = 200$ (lower curves).

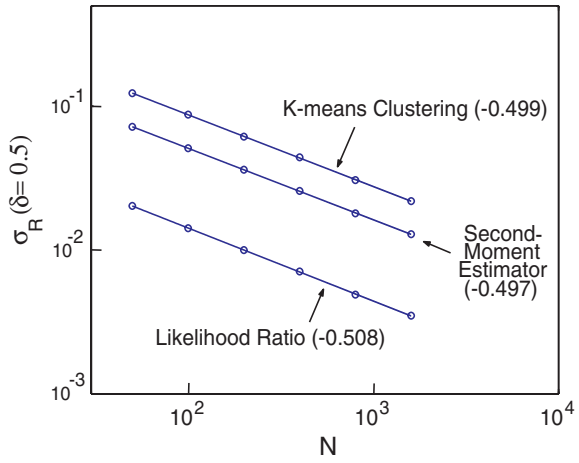

Figure 5. Standard deviation of $\hat{\delta}$ for $\delta = 0.5$ (normalized by σ) as a function of N; circles: simulation results; lines: regression; numbers in parentheses: slope of the regressed line; results are based on 10 000 runs per N for each method.

6. Discussion

The fact that the three different methods we used in this paper produce very similar results raises the possibility that these results set an upper limit on the performance of any other existing method for solving our problem. Indeed, we argue that, assuming the fourth-root law

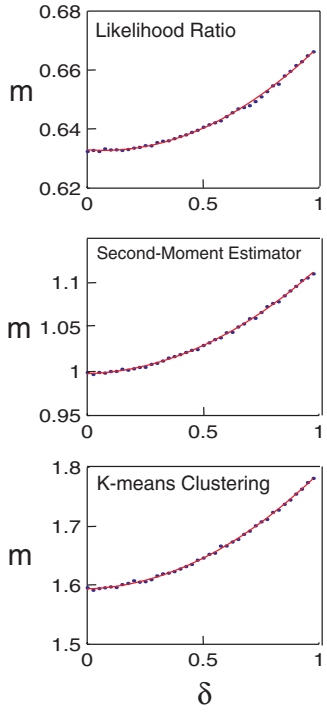

Figure 6. The δ–m data for all three methods; $N = 200$; results are based on 1000 runs per value of δ; δ is normalized by σ.

is general, the prefactors that we found for the present methods are very close to being optimal. Consider figure 2. The intersection of the N–δ lines for the mixed-set methods with that of the separate-set case happens around $N = 1$. This ensures that, for all $N > 1$, the mixed-set resolution is not better than that of the separate case, as expected. On the other hand, if the mixed-set errors were to be smaller than those found here, we would have an inconsistency because for some $N > 1$ mixing the measurements would actually reduce the error. Therefore, the performance of these two methods appears to be near-optimal. Of course, if a different method produces a different exponent β which has a larger absolute value, it can outperform the methods presented here. An example of a sub-optimal method, at least in our problem, is the use of the $J_e(2)/J_e(1)$ statistic suggested by Duda *et al* (2000). The corresponding A and β values for the N–δ data from this method are 4.8 and -0.17, respectively. These numbers indicate both a slower rate of improvement and a lower overall level of performance for all N.

The similarity between the results obtained by the three different methods that we used is parallel to results obtained by Furman and Lindsay (1994). They observe that using the method of moments has computational advantages over iterative methods, such as EM, for finding the ML-estimates, while it affects the power of the test only minimally. Moreover, they observe that using the moment estimators to initialize the ML-estimations, a small number of iterations is adequate for accurate results. This is consistent with our observation that the results from the second-moment estimator and k-means clustering methods were very close.

As discussed in section 5, all of the premises required to prove the fourth-root law hold for all the methods considered, except for the first premise in the case of the likelihood ratio method. However, our simulations show that the fourth-root law is still valid for this method.

The problem considered here is different from the hypothesis-testing problems in that, for the latter, knowledge of the test statistics is crucial in determining the confidence levels. A large body of work has addressed the null distribution of the likelihood ratio for testing the number of components in a mixture model (see, for example, Lo *et al* 2001, Ghosh and Sen 1985, Titterington *et al* 1985, Hartigan 1985a, 1985b). The main difficulty with mixture models is that regularity conditions do not hold for $-2\log\lambda$ to have its usual distribution of chi-squared with the number of degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses (McLachlan and Peel 2000, Ghosh and Sen 1985). A number of articles have addressed the properties of the likelihood ratio for mixture models where some parameters, such as mixture proportions or variances, are known (Goffinet *et al* 1992, Polymenis and Titterington 1998, Mangin *et al* 1993, Chen 1994, Chen and Cheng 1997).

In summary, we have quantified the notion that mixing noisy measurements of two distinct values of a variable not only reduces our ability to tell those values apart, but also hampers the rate of improvement gained from enlarging the size of populations, such that the resolution is proportional, no longer to $N^{-1/2}$, but to $N^{-1/4}$.

Acknowledgments

We wish to thank Dr Stanley Sawyer for his contribution to appendix B.1. This work was supported by grants from the Whitehall Foundation and the McDonnell Center for Higher Brain Function to RW.

Appendix A. Definition of the naive estimation method

This method is a one-shot approximation to the k-means clustering method. It proceeds as follows. Find the mean of all measurements. Now, compare each measurement with this mean, and classify all those points lying to the left of the mean in one cluster, and those lying to the right of the mean in a second cluster. Our estimates of the two values are the means of these two clusters:

$$\hat{\mu}_1 = \langle x_i \rangle_{X_+} \qquad \text{with} \qquad X_+ = \{x_i | x_i < \bar{x}\} \tag{A.1}$$

$$\hat{\mu}_2 = \langle x_i \rangle_{X_-} \qquad \text{with} \qquad X_- = \{x_i | x_i > \bar{x}\} \tag{A.2}$$

where $\bar{x}$ is the average over all measurements. Again, $\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1$.

This algorithm provides a very good single-pass approximation to k-means clustering. It is easy to see why this is so. If $\bar{x} = (\hat{\mu}_1 + \hat{\mu}_2)/2$, the two methods are exactly the same. The above condition, in turn, would be equivalent to saying that the number of measurements falling in X_+ and X_- is the same. This is a very reasonable approximation.

Appendix B. How does σ_R vary with N?

B.1. Naive estimation

In the following, we will prove, for the Naive estimation method, that $\sigma_R \propto N^{-1/2}$ under the condition that the mixture pdf, $f(\mathrm{x})$, can be described by a Gaussian. (This practically restricts the analysis to $\delta = 0$.) Without loss of generality, we can assume that this Gaussian is zero-mean.

To better formalize the problem, we associate two random variables, I_k and J_k, with each measurement, x_k, from the pdf, $f(x)$, with the following definition:

$$I_k = \begin{cases} 1 & : \quad x_k > \bar{x} \\ 0 & : \quad x_k \leqslant \bar{x} \end{cases} \tag{B.1}$$

and

$$J_k = \begin{cases} 0 & : \quad x_k > \bar{x} \\ 1 & : \quad x_k \leqslant \bar{x} \end{cases} \tag{B.2}$$

with $\bar{x} = \left(\sum_{k=1}^{N} x_k\right)/N$. Note that $I_k + J_k = 1$ and $I_k J_k = 0, \forall k = 1, 2, \ldots, N$. Also note that these random variables are implicitly a function of all sample points. With the above definitions, we can now re-write the expression for $\hat{\mu}_1$ and $\hat{\mu}_2$ from equations (A.1) and (A.2):

$$\hat{\mu}_1 = \frac{\sum J_k x_k}{\sum J_k} \tag{B.3}$$

$$\hat{\mu}_2 = \frac{\sum I_k x_k}{\sum I_k} \tag{B.4}$$

where all sums are over $k = 1, \ldots, N$. Our strategy is to find the means and variances and also covariances of numerators and denominators in equations (B.3) and (B.4), and use them to find the variances of the two fractions. This works as follows. If three random variables X, Y, Z are related by

$$Z = \frac{X}{Y} \tag{B.5}$$

and if the variances of X and Y are small compared to their means, then from $\delta Z = \frac{\delta X}{\bar{Y}} - \frac{\bar{X}\delta Y}{\bar{Y}^2}$, we conclude that

$$\sigma_Z^2 = \frac{\bar{X}^2}{\bar{Y}^2}\left(\frac{\sigma_X^2}{\bar{X}^2} + \frac{\sigma_Y^2}{\bar{Y}^2} - 2\frac{\mathrm{Cov}[X, Y]}{\bar{X}\bar{Y}}\right). \tag{B.6}$$

We will, therefore, be looking for $\bar{X}$, $\bar{Y}$, σ_X^2, σ_Y^2 and Cov(X, Y), where X can be $\sum I_k x_k$ or $\sum J_k x_k$, and Y can be $\sum I_k$ or $\sum J_k$. Moreover, since $\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1$:

$$\mathrm{Var}[\hat{\delta}] = \mathrm{Var}[\hat{\mu}_2] + \mathrm{Var}[\hat{\mu}_1] - 2\,\mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2] \tag{B.7}$$

then, in order to find the covariance term, we also need to find $\mathrm{E}\left[\left(\frac{\sum I_k x_k}{\sum I_k}\right)\left(\frac{\sum J_k x_k}{\sum J_k}\right)\right]$.

To begin with, consider $\sum I_k$. We first find the mean,

$$\mathrm{E}\left[\sum I_k\right] = N\mathrm{E}[I_1] = N\,\mathrm{Prob}(x_1 > \bar{x}) \tag{B.8}$$

where the first equality is due to the identical distribution of all x_k. Using the definition of $\bar{x}$

$$\mathrm{E}\left[\sum I_k\right] = N\,\mathrm{Prob}\left(y = \left(1 - \frac{1}{N}\right)x_1 + \left(-\frac{1}{N}\right)x_2 + \cdots + \left(-\frac{1}{N}\right)x_N > 0\right). \tag{B.9}$$

Now we note that y is a linear combination of Gaussian random variables, and is therefore a Gaussian itself. Since all x_k are zero-mean, then y is also zero-mean. Thus, it is distributed evenly around zero and so $\mathrm{Prob}(y > 0) = 1/2$. As a result,

$$\mathrm{E}\left[\sum I_k\right] = N/2. \tag{B.10}$$

Calculating $\mathrm{E}\left[\left(\sum I_k\right)^2\right]$ is more complicated

$$\left(\sum I_k\right)^2 = \sum I_k^2 + \sum_{m \neq n}\sum_{n} I_m I_n. \tag{B.11}$$

Thus,

$$\mathrm{E}\left[\left(\sum \mathrm{I_k}\right)^2\right] = N\mathrm{E}\left[\mathrm{I_1^2}\right] + N(N-1)\mathrm{E}[\mathrm{I_1 I_2}]. \tag{B.12}$$

Using the definition of $\mathrm{I_k}$, we note that

$$\mathrm{E}\left[\mathrm{I_1^2}\right] = \mathrm{E}[\mathrm{I_1}] = \tfrac{1}{2}. \tag{B.13}$$

On the other hand,

$$\begin{aligned}\mathrm{E}[\mathrm{I_1 I_2}] = \mathrm{Prob}\Bigg(&\mathrm{y_1} = \left(1-\frac{1}{N}\right)\mathrm{x_1} + \left(-\frac{1}{N}\right)\mathrm{x_2} + \cdots + \left(-\frac{1}{N}\right)\mathrm{x_N} > 0,\\ &\mathrm{y_2} = \left(-\frac{1}{N}\right)\mathrm{x_1} + \left(1-\frac{1}{N}\right)\mathrm{x_2} + \cdots + \left(-\frac{1}{N}\right)\mathrm{x_N} > 0\Bigg).\end{aligned} \tag{B.14}$$

Again, it is clear that $(\mathrm{y_1}, \mathrm{y_2})$ are jointly Gaussian. Therefore, to calculate $\mathrm{E}[\mathrm{I_1 I_2}]$, we need to find P, the covariance matrix for $(\mathrm{y_1}, \mathrm{y_2})$. Since $\mathrm{Cov}(\mathrm{x_i}, \mathrm{x_j}) = \delta_{\mathrm{i,j}}$, it is easy to verify that

$$\mathrm{P} = \begin{pmatrix} \mathrm{E}\left[\mathrm{y_1^2}\right] & \mathrm{E}[\mathrm{y_1 y_2}] \\ \mathrm{E}[\mathrm{y_1 y_2}] & \mathrm{E}\left[\mathrm{y_2^2}\right] \end{pmatrix} = \sigma^2 \begin{pmatrix} (N-1)/N & -1/N \\ -1/N & (N-1)/N \end{pmatrix}. \tag{B.15}$$

Equation (B.14) will now translate into

$$E[\mathrm{I_1 I_2}] = \int_0^\infty \int_0^\infty f(y_1', y_2')\, \mathrm{d}y_1'\, \mathrm{d}y_2' \tag{B.16}$$

with

$$f(\mathrm{y_1}, \mathrm{y_2}) = \frac{1}{2\pi |\mathrm{P}|^{1/2}}\, \mathrm{e}^{-\frac{1}{2}\vec{\mathrm{y}}^T P^{-1}\vec{\mathrm{y}}} \tag{B.17}$$

where $\vec{\mathrm{y}} = [\mathrm{y_1}, \mathrm{y_2}]^T$, and P is defined in equation (B.15). First, we note that

$$\mathrm{P}^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} (N-1)/(N-2) & 1/(N-2) \\ 1/(N-2) & (N-1)/(N-2) \end{pmatrix} \tag{B.18}$$

$$|P| = \sigma^4 \frac{N-2}{N}. \tag{B.19}$$

Therefore,

$$\begin{aligned}\mathrm{E}[\mathrm{I_1 I_2}] = \frac{1}{2\pi\sigma^2}\sqrt{\frac{N}{N-2}} \int_0^{+\infty}\int_0^{+\infty} \\ \times \exp\left(\frac{-1}{2\sigma^2(N-2)}\left[(N-1)y_1'^2 + (N-1)y_2'^2 + 2y_1'y_2'\right]\right) \mathrm{d}y_1'\, \mathrm{d}y_2'.\end{aligned} \tag{B.20}$$

But for large N,

$$\begin{aligned}\exp\left(\frac{-1}{2\sigma^2(N-2)}\left[(N-1)y_1'^2 + (N-1)y_2'^2 + 2y_1'y_2'\right]\right) &\approx \exp\left(-\frac{y_1'^2 + y_2'^2 + \frac{(y_1'+y_2'^2)}{N}}{2\sigma^2}\right) \\ &\approx \exp\left(-\frac{y_1'^2 + y_2'^2}{2\sigma^2}\right)\left(1 - \frac{(y_1' + y_2'^2)}{2\sigma^2}\frac{1}{N}\right).\end{aligned} \tag{B.21}$$

Also, note that

$$\sqrt{\frac{N}{N-2}} = \left(1 - \frac{2}{N}\right)^{-1/2} \approx 1 + \frac{1}{N}. \tag{B.22}$$

Therefore, combining equations (B.20), (B.21) and (B.22),

$$\mathrm{E}[\mathrm{I}_1\mathrm{I}_2] \approx A + \frac{B}{N} \tag{B.23}$$

with

$$A = \frac{1}{2\pi\sigma^2}\int_0^{+\infty}\int_0^{+\infty}\exp\left(-\frac{y_1'^2 + y_1'^2}{2\sigma^2}\right)\mathrm{d}y_1'\,\mathrm{d}y_2' \tag{B.24}$$

$$\begin{aligned} B = \frac{1}{2\pi\sigma^2}\Bigg[&\int_0^{+\infty}\int_0^{+\infty}\exp\left(-\frac{y_1'^2 + y_1'^2}{2\sigma^2}\right)\mathrm{d}y_1'\,\mathrm{d}y_2' \\ &- \frac{1}{2\sigma^2}\int_0^{+\infty}\int_0^{+\infty}(y_1' + y_2')^2\exp\left(-\frac{y_1'^2 + y_2'^2}{2\sigma^2}\right)\mathrm{d}y_1'\,\mathrm{d}y_2'\Bigg]. \end{aligned} \tag{B.25}$$

The above integrals are elementary, and the result is

$$\mathrm{E}[\mathrm{I}_1\mathrm{I}_2] \approx \frac{1}{4}\left(1 - \frac{2}{\pi}\frac{1}{N}\right). \tag{B.26}$$

From equations (C.6), (B.13) and (B.26), we conclude

$$\mathrm{E}\left[\left(\sum \mathrm{I}_\mathrm{k}\right)^2\right] \approx \frac{N}{2} + \frac{N^2}{4}\left[1 - \left(1 + \frac{2}{\pi}\right)\frac{1}{N}\right]. \tag{B.27}$$

Therefore,

$$\begin{aligned} \mathrm{Var}\left[\sum \mathrm{I}_\mathrm{k}\right] &\approx \frac{N}{2} + \frac{N^2}{4}\left[1 - \left(1 + \frac{2}{\pi}\right)\frac{1}{N}\right] - \frac{N^2}{4} \\ &= \frac{N}{4}\left(1 - \frac{2}{\pi}\right). \end{aligned} \tag{B.28}$$

In summary, we found that

$$\mathrm{E}\left[\sum \mathrm{I}_\mathrm{k}\right] = \frac{N}{2} \qquad \text{and} \qquad \mathrm{Var}\left[\sum \mathrm{I}_\mathrm{k}\right] \approx \frac{N}{4}\left(1 - \frac{2}{\pi}\right). \tag{B.29}$$

Next, we consider $\sum \mathrm{I}_\mathrm{k}\mathrm{x}_\mathrm{k}$,

$$\mathrm{E}\left[\sum \mathrm{I}_\mathrm{k}\mathrm{x}_\mathrm{k}\right] = N\mathrm{E}[\mathrm{I}_1\mathrm{x}_1] = N\int_{y_2'=0}^{\infty}\int_{y_1'=-\infty}^{+\infty} y_1' f(y_1', y_2')\,\mathrm{d}y_1'\,\mathrm{d}y_2' \tag{B.30}$$

where $f(\mathrm{y}_1, \mathrm{y}_2)$ is the joint pdf for $\mathrm{y}_1 = \mathrm{x}_1$ and $\mathrm{y}_2 = (1 - 1/N)\mathrm{x}_1 + (-1/N)\mathrm{x}_2 + \cdots + (-1/N)\mathrm{x}_\mathrm{N}$ (note the limits of the integrals). The function f is described by (B.17), with P calculated to be

$$\begin{aligned} &\mathrm{P} = \sigma^2\begin{pmatrix} 1 & (N-1)/N \\ (N-1)/N & (N-1)/N \end{pmatrix} \quad \Longrightarrow \\ &\mathrm{P}^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} N & -N \\ -N & N^2/(N-1) \end{pmatrix} \qquad \text{and} \qquad |\mathrm{P}| = \sigma^4\frac{N-1}{N^2}. \end{aligned} \tag{B.31}$$

Therefore,

$$\begin{aligned} f(\mathrm{y}_1, \mathrm{y}_2) &= \frac{N}{2\pi\sigma^2\sqrt{N-1}}\exp\left(-\frac{N}{2\sigma^2}\left[\mathrm{y}_1^2 + \frac{\mathrm{N}}{\mathrm{N}-1}\mathrm{y}_2^2 - 2\mathrm{y}_1\mathrm{y}_2\right]\right) \\ &\approx \frac{N^{1/2}}{2\pi\sigma^2}\left(1 + \frac{1}{2}\frac{1}{N}\right)\exp\left(-\frac{N}{2\sigma^2}(y_1' - y_2'^2)\right)\exp\left(-\frac{y_2'^2}{2\sigma^2}\left(1 + \frac{1}{N}\right)\right). \end{aligned} \tag{B.32}$$

As a result,

$$\begin{aligned}\mathrm{E}\left[\sum \mathrm{I_k x_k}\right] &\approx \frac{N^{3/2}}{2\pi\sigma^2}\left(1+\frac{1}{2}\frac{1}{N}\right)\int_{y_2'=0}^{+\infty}\mathrm{d}y_2'\exp\left(-\frac{y_2'^2}{2\sigma^2}\left(1+\frac{1}{N}\right)\right)\\ &\quad\times\int_{y_1'=-\infty}^{+\infty}\mathrm{d}y_1'\,y_1'\exp\left(-\frac{N}{2\sigma^2}(y_1'-y_2')^2\right)\\ &= \frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right). \end{aligned} \tag{B.33}$$

To find the variance, we need $\mathrm{E}\left[\left(\sum \mathrm{I_k x_k}\right)^2\right]$:

$$\begin{aligned}\mathrm{E}\left[\left(\sum \mathrm{I_k x_k}\right)^2\right] &= \mathrm{E}\left[\sum \mathrm{I_k x_k^2}\right] + \mathrm{E}\left[\sum_{\mathrm{m}\neq\mathrm{n}} \mathrm{I_m I_n x_m x_n}\right]\\ &= N\mathrm{E}\left[\mathrm{I_1 x_1^2}\right] + N(N-1)\mathrm{E}[\mathrm{I_1 I_2 x_1 x_2}]. \end{aligned} \tag{B.34}$$

To find the first term, we can use the set-up of the previous case to find

$$\begin{aligned}E\left[\sum \mathrm{I_k x_k^2}\right] &\approx \frac{N^{3/2}}{2\pi\sigma^2}\left(1+\frac{1}{2}\frac{1}{N}\right)\int_{y_2'=0}^{+\infty}\mathrm{d}y_2'\exp\left(-\frac{y_2'^2}{2\sigma^2}\left(1+\frac{1}{N}\right)\right)\\ &\quad\times\int_{y_1'=-\infty}^{+\infty}\mathrm{d}y_1'\,y_1'^2\exp\left(-\frac{N}{2\sigma^2}(y_1'-y_2')^2\right). \end{aligned} \tag{B.35}$$

Using the fact that $y_1'^2 = (y_1'-y_2')^2 - 2y_2'(y_1'-y_2') + y_2'^2$, we arrive at the following:

$$\begin{aligned}\mathrm{E}\left[\sum \mathrm{I_k x_k^2}\right] &\approx \frac{N^{3/2}}{2\pi\sigma^2}\left(1+\frac{1}{2}\frac{1}{N}\right)\int_{y_2'=0}^{+\infty}\mathrm{d}y_2'\exp\left(-\frac{y_2'^2}{2\sigma^2}\left(1+\frac{1}{N}\right)\right)\\ &\quad\times\left[\left(\frac{\sigma}{\sqrt{N}}\right)^3\sqrt{2\pi}+y_2'^2\frac{\sigma}{\sqrt{N}}\sqrt{2\pi}\right]\\ &\approx \frac{N}{\sigma\sqrt{2\pi}}\left(1+\frac{1}{2}\frac{1}{N}\right)\int_{y_2'=0}^{+\infty}\mathrm{d}y_2'\exp\left(-\frac{y_2'^2}{2\sigma^2}\left(1+\frac{1}{N}\right)\right)\left(\frac{\sigma^2}{N}+y_2'^2\right)\\ &\approx \frac{N\sigma^2}{2}. \end{aligned} \tag{B.36}$$

For $\mathrm{E}[\mathrm{I_1 I_2 x_1 x_2}]$, we define the variables

$$\begin{aligned}\mathrm{z_1} &= \mathrm{x_1}\\ \mathrm{z_2} &= \left(1-\frac{1}{N}\right)\mathrm{x_1}+\left(-\frac{1}{N}\right)\mathrm{x_2}+\cdots+\left(-\frac{1}{N}\right)\mathrm{x_N}\\ \mathrm{z_3} &= \left(-\frac{1}{N}\right)\mathrm{x_1}+\left(1-\frac{1}{N}\right)\mathrm{x_2}+\cdots+\left(-\frac{1}{N}\right)\mathrm{x_N}. \end{aligned} \tag{B.37}$$

With the above definitions, $\mathrm{x_2} = \mathrm{z_1} - \mathrm{z_2} + \mathrm{z_3}$. Now we note that

$$\mathrm{E}[\mathrm{I_1 I_2 x_1 x_2}] = \int_{z_3'=0}^{+\infty}\int_{z_2'=0}^{+\infty}\int_{z_1'=-\infty}^{+\infty} z_1'(z_1'+z_3'-z_2')\,f(z_1', z_2', z_3')\,\mathrm{d}z_1'\,\mathrm{d}z_2'\,\mathrm{d}z_3'. \tag{B.38}$$

The covariance matrix for $(\mathrm{z_1}, \mathrm{z_2}, \mathrm{z_3})$ is

$$\mathrm{P} = \sigma^2\begin{pmatrix} 1 & (N-1)/N & -1/N\\ (N-1)/N & (N-1)/N & -1/N\\ -1/N & -1/N & (N-1)/N\end{pmatrix}. \tag{B.39}$$

So,

$$\mathrm{P}^{-1} = \sigma^2 \begin{pmatrix} N & -N & 0 \\ -N & (N^2 - N - 1)/(N-2) & 1/(N-2) \\ 0 & 1/(N-2) & (N-1)/(N-2) \end{pmatrix} \tag{B.40}$$

$$|\mathrm{P}| = \frac{N-2}{N^2}. \tag{B.41}$$

Therefore,

$$\begin{aligned} \mathrm{E}[\mathrm{I}_1\mathrm{I}_2\mathrm{x}_1\mathrm{x}_2] = {} & \frac{N}{(2\pi)^{3/2}(N-2)^{1/2}\sigma^3} \int_{z_3'=0}^{+\infty}\int_{z_2'=0}^{+\infty}\int_{z_1'=-\infty}^{+\infty} z_1'(z_1' + z_3' - z_2') \\ & \times \exp\left(-\frac{1}{2\sigma^2}\left\{N z_1'^2 + \frac{N^2-N-1}{N-2} z_2'^2 + \frac{N-1}{N-2} z_3'^2 \right.\right. \\ & \left.\left. - 2N z_1' z_2' + \frac{2}{N-2} z_2' z_3'\right\}\right) \mathrm{d}z_1'\,\mathrm{d}z_2'\,\mathrm{d}z_3'. \end{aligned} \tag{B.42}$$

Now we note that

$$\begin{aligned} & N z_1'^2 + \frac{N^2-N-1}{N-2} z_2'^2 + \frac{N-1}{N-2} z_3'^2 - 2N z_1' z_2' + \frac{2}{N-2} z_2' z_3' \\ & \quad = N(z_1' - z_2')^2 + \frac{1}{N-2}\big[(N-1)\big(z_2'^2 + z_3'^2\big) + 2z_2' z_3'\big]. \end{aligned} \tag{B.43}$$

Therefore,

$$\begin{aligned} \mathrm{E}\left[\sum_{\mathrm{m}\neq\mathrm{n}} \mathrm{I}_\mathrm{m}\mathrm{I}_\mathrm{n}\mathrm{x}_\mathrm{m}\mathrm{x}_\mathrm{n}\right] = {} & N(N-1)\frac{N}{(2\pi)^{3/2}(N-2)^{1/2}\sigma^3} \\ & \times \int_0^{+\infty}\int_0^{+\infty} \mathrm{d}z_2'\,\mathrm{d}z_3' \exp\left(-\frac{1}{2(N-2)\sigma^2}\big[(N-1)\big(z_2'^2 + z_3'^2\big) + 2z_2' z_3'\big]\right) \\ & \times \int_{-\infty}^{+\infty} z_1'(z_1' - z_2' + z_3') \exp\left(-\frac{N}{2\sigma^2}(z_1' - z_2')^2\right) \mathrm{d}z_1'. \end{aligned} \tag{B.44}$$

Next, we note that

$$z_1'(z_1' - z_2' + z_3') = (z_1' - z_2')^2 + (z_2' + z_3')(z_1' - z_2') + z_2' z_3'. \tag{B.45}$$

Therefore,

$$\begin{aligned} \mathrm{E}\left[\sum_{\mathrm{m}\neq\mathrm{n}} \mathrm{I}_\mathrm{m}\mathrm{I}_\mathrm{n}\mathrm{x}_\mathrm{m}\mathrm{x}_\mathrm{n}\right] = {} & N(N-1)\frac{N}{(2\pi)^{3/2}(N-2)^{1/2}\sigma^3} \\ & \times \int_0^{+\infty}\int_0^{+\infty} \mathrm{d}z_2'\,\mathrm{d}z_3' \exp\left(-\frac{1}{2(N-2)\sigma^2}\big[(N-1)\big(z_2'^2 + z_3'^2\big) + 2z_2' z_3'\big]\right) \\ & \times \left[\left(\frac{\sigma}{\sqrt{N}}\right)^3 \sqrt{2\pi} + \left(\frac{\sigma}{\sqrt{N}}\sqrt{2\pi}\right) z_2' z_3'\right]. \end{aligned} \tag{B.46}$$

Since

$$\frac{N-1}{\sqrt{N-2}} = \sqrt{N} + O\big(N^{-\frac{3}{2}}\big) \tag{B.47}$$

and

$$\frac{1}{(N-2)}\big[(N-1)\big(z_2'^2 + z_3'^2\big) + 2z_2' z_3'\big] = z_2'^2 + z_3'^2 + \frac{(z_2' + z_3')^2}{N} + O\left(\frac{1}{N^2}\right) \tag{B.48}$$

we have

$$\exp\left(-\frac{1}{2(N-2)\sigma^2}\left[(N-1)\left(z_2'^2+z_3'^2\right)+2z_2'z_3'\right]\right)\approx\exp\left(-\frac{z_2'^2+z_3'^2}{2\sigma^2}\right)\left[1-\frac{(z_2'+z_3')^2}{2\sigma^2}\frac{1}{N}\right]. \tag{B.49}$$

Putting everything together, we find

$$\begin{aligned}\mathrm{E}\left[\sum_{m\neq n}\mathrm{I_m I_n x_m x_n}\right]&\approx\frac{N^2}{2\pi\sigma^2}\int_0^{+\infty}\int_0^{+\infty}\mathrm{d}z_2'\,\mathrm{d}z_3'\exp\left(-\frac{z_2'^2+z_3'^2}{2\sigma^2}\right)\\&\quad\times\left[1-\frac{(z_2'+z_3')^2}{2\sigma^2}\frac{1}{N}\right]\left(\frac{\sigma^2}{N}+z_2'z_3'\right)\\&\approx\frac{N^2}{2\pi\sigma^2}\int_0^{+\infty}\int_0^{+\infty}\mathrm{d}z_2'\,\mathrm{d}z_3'\exp\left(-\frac{z_2'^2+z_3'^2}{2\sigma^2}\right)\\&\quad\times\left[z_2'z_3'+\left(\sigma^2-\frac{z_2'z_3'(z_2'+z_3')^2}{2\sigma^2}\right)\frac{1}{N}\right]\\&\approx\frac{N^2\sigma^2}{2\pi}\left(1-\frac{2}{N}\right).\end{aligned} \tag{B.50}$$

From equations (B.34), (B.36) and (B.50), we conclude that

$$\mathrm{E}\left[\left(\sum\mathrm{I_k x_k}\right)^2\right]=\frac{\mathrm{N}\sigma^2}{2}+\frac{\mathrm{N}^2\sigma^2}{2\pi}\left(1-\frac{2}{\mathrm{N}}\right). \tag{B.51}$$

Since $\mathrm{Var}\left[\sum\mathrm{I_k x_k}\right]=\mathrm{E}\left[\left(\sum\mathrm{I_k x_k}\right)^2\right]-(\mathrm{E}[\mathrm{I_k x_k}])^2$, and from equations (B.33) and (B.51), we find

$$\begin{aligned}\mathrm{Var}\left[\sum\mathrm{I_k x_k}\right]&\approx\frac{N\sigma^2}{2}+\frac{N^2\sigma^2}{2\pi}\left(1-\frac{2}{N}\right)-\left[\frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right)\right]^2\\&\approx\frac{N\sigma^2}{2}\left(1-\frac{1}{\pi}\right).\end{aligned} \tag{B.52}$$

In summary,

$$\mathrm{E}\left[\sum\mathrm{I_k x_k}\right]\approx\frac{N\sigma}{\sqrt{2\pi}}\qquad\text{and}\qquad\mathrm{Var}\left[\sum\mathrm{I_k x_k}\right]\approx\frac{N\sigma^2}{2}\left(1-\frac{1}{\pi}\right). \tag{B.53}$$

Next, we consider $\mathrm{Cov}\left[\sum\mathrm{I_k},\sum\mathrm{I_k x_k}\right]$

$$\mathrm{Cov}\left[\sum\mathrm{I_k},\sum\mathrm{I_k x_k}\right]=\mathrm{E}\left[\sum\mathrm{I_k}\sum\mathrm{I_k x_k}\right]-\mathrm{E}\left[\sum\mathrm{I_k}\right]\mathrm{E}\left[\sum\mathrm{I_k x_k}\right]. \tag{B.54}$$

But

$$\begin{aligned}\mathrm{E}\left[\sum\mathrm{I_k}\sum\mathrm{I_k x_k}\right]&=\mathrm{E}\left[\sum\mathrm{I_k x_k}\right]+\mathrm{E}\left[\sum_{m,n\neq m}\mathrm{I_m I_n x_n}\right]\\&\approx\frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right)+N(N-1)\mathrm{E}[\mathrm{I_1 I_2 x_1}].\end{aligned} \tag{B.55}$$

The same change of variables that were introduced in equation (C.6) can be used in this case as well, leading us to the following:

$$\begin{aligned}\mathrm{E}[\mathrm{I_1 I_2 x_1}]=\frac{N}{(2\pi)^{3/2}(N-2)^{1/2}\sigma^3}\int_{z_3'=0}^{+\infty}\int_{z_2'=0}^{+\infty}\int_{z_1'=-\infty}^{+\infty}z_1'\exp\left(-\frac{1}{2\sigma^2}\left\{Nz_1'^2+\frac{N^2-N-1}{N-2}z_2'^2\right.\right.\\\left.\left.+\frac{N-1}{N-2}z_3'^2-2Nz_1'z_2'+\frac{2}{N-2}z_2'z_3'\right\}\right)\mathrm{d}z_1'\,\mathrm{d}z_2'\,\mathrm{d}z_3'.\end{aligned} \tag{B.56}$$

Again, very similar to equation (B.71)

$$\mathrm{E}\left[\sum_{m\neq n} I_m I_n x_n\right] = N(N-1)\frac{N}{(2\pi)^{3/2}(N-2)^{1/2}\sigma^3} \times \int_0^{+\infty}\int_0^{+\infty} dz_2'\, dz_3' \exp\left(-\frac{1}{2(N-2)\sigma^2}\left[(N-1)\left(z_2'^2+z_3'^2\right)+2z_2'z_3'\right]\right) \times \int_{-\infty}^{+\infty} z_1' \exp\left(-\frac{N}{2\sigma^2}(z_1'-z_2'^2)\right) dz_1' \tag{B.57}$$

$$\approx \frac{N^2\sigma}{2\sqrt{2\pi}}\left(1-\frac{5}{2}\frac{1}{N}\right). \tag{B.58}$$

As a result,

$$\mathrm{E}\left[\sum I_k \sum I_k x_k\right] \approx \frac{N^2\sigma}{2\sqrt{2\pi}}\left(1-\frac{5}{2}\frac{1}{N}\right)+\frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right) \tag{B.59}$$

which, in turn, means

$$\begin{aligned}\mathrm{Cov}\left[\sum I_k, \sum I_k x_k\right] &\approx \frac{N^2\sigma}{2\sqrt{2\pi}}\left(1-\frac{5}{2}\frac{1}{N}\right)+\frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right)-\left[\frac{N}{2}\right]\left[\frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right)\right] \\ &\approx \frac{N\sigma}{2\sqrt{2\pi}}\left[N-\frac{5}{2}+2-\frac{1}{N}-N+\frac{1}{2}\right] \\ &= O(1).\end{aligned} \tag{B.60}$$

Combining equations (B.29), (B.53) and (B.60), we can calculate $\mathrm{Var}[\hat{\mu}_2]$

$$\begin{aligned}\mathrm{Var}[\hat{\mu}_2] &= \mathrm{Var}\left[\frac{\sum I_k x_k}{\sum I_k}\right] \\ &= \frac{\left(\mathrm{E}\left[\sum I_k x_k\right]\right)^2}{\left(\mathrm{E}\left[\sum I_k\right]\right)^2}\left(\frac{\mathrm{Var}\left[\sum I_k x_k\right]}{\left(\mathrm{E}\left[\sum I_k x_k\right]\right)^2}+\frac{\mathrm{Var}\left[\sum I_k\right]}{\left(\mathrm{E}\left[\sum I_k\right]\right)^2}-2\frac{\mathrm{Cov}\left[\sum I_k x_k, \sum I_k\right]}{\mathrm{E}\left[\sum I_k x_k\right]\mathrm{E}\left[\sum I_k\right]}\right) \\ &= \frac{\left(\frac{N\sigma}{\sqrt{2\pi}}\right)^2}{\left(\frac{N}{2}\right)^2}\left[\frac{\frac{N}{4}\left(1-\frac{2}{\pi}\right)}{\left(\frac{N}{2}\right)^2}+\frac{\frac{N\sigma^2}{2}\left(1-\frac{1}{\pi}\right)}{\left(\frac{N\sigma}{\sqrt{2\pi}}\right)^2}+O\left(\frac{1}{N^2}\right)\right] \\ &\approx \frac{2\sigma^2}{N}\left(1-\frac{2}{\pi^2}\right).\end{aligned} \tag{B.61}$$

From symmetry considerations, the same result is valid for $\hat{\mu}_1$.

Finally, we need to find $\mathrm{Var}[\hat{\mu}_2-\hat{\mu}_1]$ for which we have to find $\mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2]$:

$$\mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2] = \mathrm{E}[\hat{\mu}_1\hat{\mu}_2] - \mathrm{E}[\hat{\mu}_1]\mathrm{E}[\hat{\mu}_2]. \tag{B.62}$$

In order to calculate the above expectations, we use the following theorem which holds when standard deviations are much smaller than their corresponding means

$$\mathrm{E}\left[\frac{\mathrm{X}}{\mathrm{Y}}\right] = \frac{\bar{\mathrm{X}}}{\bar{\mathrm{Y}}}\left(1+\frac{\mathrm{Var}[\mathrm{X}]}{\bar{\mathrm{X}}^2}-\frac{\mathrm{Cov}[\mathrm{X},\mathrm{Y}]}{\bar{\mathrm{X}}\bar{\mathrm{Y}}}\right). \tag{B.63}$$

We therefore have

$$\mathrm{E}[\hat{\mu}_2] = -\mathrm{E}[\hat{\mu}_1] = \frac{\mathrm{E}\left[\sum I_k x_k\right]}{\mathrm{E}\left[\sum I_k\right]}\left(1+\frac{\mathrm{Var}\left[\sum I_k\right]}{\left(\mathrm{E}\left[\sum I_k\right]\right)^2}-\frac{\mathrm{Cov}\left[\sum I_k, \sum I_k x_k\right]}{\mathrm{E}\left[\sum I_k\right]\mathrm{E}\left[\sum I_k x_k\right]}\right) \tag{B.64}$$

$$\begin{aligned}&= \frac{\frac{N\sigma}{\sqrt{2\pi}}\left(1-\frac{1}{2}\frac{1}{N}\right)}{\frac{N}{2}}\left(1+\frac{\frac{N}{4}\left(1-\frac{2}{\pi}\right)}{\frac{N^2}{4}}+O\left(\frac{1}{N^2}\right)\right) \\ &= \sigma\sqrt{\frac{2}{\pi}}\left(1+\left(\frac{1}{2}-\frac{2}{\pi}\right)\frac{1}{N}\right).\end{aligned} \tag{B.65}$$

As a result,

$$\mathrm{E}[\hat{\mu}_1]\mathrm{E}[\hat{\mu}_2] = -\frac{2\sigma^2}{\pi}\left(1 + \left(1 - \frac{4}{\pi}\right)\frac{1}{N}\right). \tag{B.66}$$

The last piece we need is

$$\begin{aligned}\mathrm{E}[\hat{\mu}_1\hat{\mu}_2] &= \mathrm{E}\left[\frac{\sum \mathrm{I_k x_k} \sum \mathrm{J_k x_k}}{\sum \mathrm{I_k} \sum \mathrm{J_k}}\right] = \mathrm{E}\left[\frac{\sum \mathrm{I_m J_n x_m x_n}}{\sum \mathrm{I_m J_n}}\right] \\ &= \frac{\mathrm{E}\left[\sum \mathrm{I_m J_n x_m x_n}\right]}{\mathrm{E}\left[\sum \mathrm{I_m J_n}\right]}\left(1 + \frac{\mathrm{Var}\left[\sum \mathrm{I_m J_n}\right]}{\left(\mathrm{E}\left[\sum \mathrm{I_m J_n}\right]\right)^2} - \frac{\mathrm{Cov}\left[\sum \mathrm{I_m J_n x_m x_n}, \sum \mathrm{I_m J_n}\right]}{\mathrm{E}\left[\sum \mathrm{I_m J_n x_m x_n}\right]\mathrm{E}\left[\sum \mathrm{I_m J_n}\right]}\right).\end{aligned} \tag{B.67}$$

Calculations very similar to those above show that the last two terms in parentheses do not contribute to the order of magnitude we are looking for. It can also be shown that

$$\mathrm{E}\left[\sum \mathrm{I_m J_n}\right] = \frac{\mathrm{N}^2}{4}\left(1 + \left(\frac{2}{\pi} - 1\right)\frac{1}{\mathrm{N}}\right) \tag{B.68}$$

$$\mathrm{E}\left[\sum \mathrm{I_m J_n x_m x_n}\right] = -\frac{N^2\sigma^2}{2\pi}\left(1 - \frac{2}{N}\right). \tag{B.69}$$

Putting everything together, we find

$$\mathrm{E}[\hat{\mu}_1\hat{\mu}_2] = \frac{\mathrm{E}\left[\sum \mathrm{I_m J_n x_m x_n}\right]}{\mathrm{E}\left[\sum \mathrm{I_m J_n}\right]} = -\frac{2\sigma^2}{\pi}\left(1 - \left(1 + \frac{2}{\pi}\right)\frac{1}{N}\right). \tag{B.70}$$

Using equations (B.62), (B.66) and (B.70), we find

$$\begin{aligned}\mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2] &= -\frac{2\sigma^2}{\pi}\left(1 - \left(1 + \frac{2}{\pi}\right)\frac{1}{N}\right) + \frac{2\sigma^2}{\pi}\left(1 + \left(1 - \frac{4}{\pi}\right)\frac{1}{N}\right) \\ &= \frac{4\sigma^2}{N}\left(\frac{1}{\pi} - \frac{1}{\pi^2}\right).\end{aligned} \tag{B.71}$$

Now we can calculate the variance of $\hat{\mu}_2 - \hat{\mu}_1$

$$\begin{aligned}\mathrm{Var}[\hat{\mu}_2 - \hat{\mu}_1] &= \mathrm{Var}[\hat{\mu}_1] + \mathrm{Var}[\hat{\mu}_2] - 2\,\mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2] \\ &= 2(\mathrm{Var}[\hat{\mu}_2] - \mathrm{Cov}[\hat{\mu}_1, \hat{\mu}_2]).\end{aligned} \tag{B.72}$$

Using equations (B.61) and (B.71), we find

$$\begin{aligned}\mathrm{Var}[\hat{\mu}_2 - \hat{\mu}_1] &= 2\left(\frac{2\sigma^2}{N}\left(1 - \frac{2}{\pi^2}\right) - \frac{4\sigma^2}{N}\left(\frac{1}{\pi} - \frac{1}{\pi^2}\right)\right) \\ &= \frac{4\sigma^2}{N}\left(1 - \frac{2}{\pi}\right).\end{aligned} \tag{B.73}$$

Obviously,

$$\hat{\sigma} = \frac{2\sigma\sqrt{1 - 2/\pi}}{\sqrt{N}}. \tag{B.74}$$

This concludes the proof that $\hat{\sigma}$ is proportional to $N^{-1/2}$.

B.2. Method of moments

We first prove that, for large N,

$$\mathrm{Var}[V_2] \propto \frac{1}{N} \tag{B.75}$$

with V_2 defined in equation (5). Using the fact that $\bar{x} = (x_1 + x_2 + \cdots + x_N)/N$, we can re-arrange equation (5) to find that

$$V_2 = \left(\frac{N-1}{N^2}\right)\sum_i x_i^2 + \left(\frac{-1}{N^2}\right)\sum_i \sum_{j\neq i} x_i x_j. \tag{B.76}$$

Therefore,

$$\mathrm{Var}[V_2] = \left(\frac{N-1}{N^2}\right)^2 \mathrm{Var}\left[\sum x_i^2\right] + \frac{1}{N^4}\mathrm{Var}\left[\sum_i \sum_{j\neq i} x_i x_j\right] - \frac{N-1}{N^4}\mathrm{Cov}\left[\sum x_i^2, \sum_i \sum_{j\neq i} x_i x_j\right]. \tag{B.77}$$

As for the first term,

$$\mathrm{Var}\left[\sum_i x_i^2\right] = N\,\mathrm{Var}\left[x_1^2\right] + N(N-1)\mathrm{Cov}\left[x_1^2, x_2^2\right]. \tag{B.78}$$

But since x_i are independently drawn samples, the second term in the above is zero and we have

$$\begin{aligned}\mathrm{Var}\left[\sum_i x_i^2\right] &= N\,\mathrm{Var}\left[x_1^2\right] \\ &= O(N).\end{aligned} \tag{B.79}$$

Next consider the second term in equation (B.77):

$$\begin{aligned}\mathrm{Var}\left[\sum_i \sum_{j\neq i} x_i x_j\right] &= 2N(N-1)\mathrm{Var}[x_1x_2] + N(N-1)(N-2)(N-3)\mathrm{Cov}[x_1x_2, x_3x_4] \\ &\quad + 4N(N-1)(N-2)\mathrm{Cov}[x_1x_2, x_1x_3] \\ &\simeq 4N(N-1)(N-2)\mathrm{Cov}[x_1x_2, x_1x_3] \\ &= O(N^3),\end{aligned} \tag{B.80}$$

while the third line is again due to the fact that samples are independent. Finally, the last term in equation (B.77) is

$$\begin{aligned}\mathrm{Cov}\left[\sum x_i^2, \sum_i \sum_{j\neq i} x_i x_j\right] &= N(N-1)(N-2)\mathrm{Cov}\left[x_1^2, x_2x_3\right] + 2N(N-1)\mathrm{Cov}\left[x_1^2, x_1x_2\right] \\ &= 2N(N-1)\mathrm{Cov}\left[x_1^2, x_1x_2\right] \\ &= O(N^2).\end{aligned} \tag{B.81}$$

Putting everything together

$$\mathrm{Var}[V_2] = \left(\frac{N-1}{N^2}\right)^2 O(N) + \frac{1}{N^4}O(N^3) - \frac{N-1}{N^4}O(N^2). \tag{B.82}$$

It is now clear that

$$\mathrm{Var}[V_2] \propto \frac{1}{N} \tag{B.83}$$

for large N.

Note that the above implies that the fluctuation of V_2 around its mean will become smaller and smaller as N gets larger. In other words,

$$\frac{\mathrm{Var}[V_2]}{\mathrm{E}[V_2]} \to 0 \qquad N \to \infty. \tag{B.84}$$

Therefore, the following is true (most of the time) for large N:

$$V_2 = \mathrm{E}[V_2] + \varepsilon \qquad \varepsilon \ll 1. \tag{B.85}$$

As a result,

$$V_2^{1/2} = (\mathrm{E}[V_2])^{1/2} + \frac{1}{2}\frac{\varepsilon}{(\mathrm{E}[V_2])^{1/2}} \tag{B.86}$$

which, in turn, means

$$\mathrm{Var}\left[V_2^{1/2}\right] = \frac{1}{4\mathrm{E}[V_2]}\mathrm{Var}[V_2]. \tag{B.87}$$

Combining equations (B.83) and (B.87), and noting that $E[V_2]$ is independent of N, results in

$$\mathrm{Var}\left[V_2^{1/2}\right] \propto \frac{1}{N}. \tag{B.88}$$

Appendix C. Quadratic shape of δ–m curve

If we denote the noise pdf for an individual measurement by $F(\mathrm{x})$, then the mixture pdf $f(\mathrm{x}; \mu_1, \mu_2)$ will have the following form:

$$f(\mathrm{x}; \mu_1, \mu_2) = \tfrac{1}{2}[F(\mathrm{x} - \mu_1) + F(\mathrm{x} - \mu_2)]. \tag{C.1}$$

A natural assumption to make about random noise is that it has zero mean. Therefore, we can state the following about $F(\mathrm{x})$:

$$\int_{-\infty}^{+\infty} F(u')\,\mathrm{d}u' = 1 \tag{C.2}$$

$$\int_{-\infty}^{+\infty} u' F(u')\,\mathrm{d}u' = 0. \tag{C.3}$$

C.1. Naive estimation method

As discussed above, the first step in the naive estimation is finding the global mean

$$\bar{x} = \int_{-\infty}^{+\infty} x' f(x')\,\mathrm{d}x' = \frac{\mu_1 + \mu_2}{2} \tag{C.4}$$

where the second equality results from equation (C.1). Next, we estimate μ_1 and μ_2 by averaging over all measurements to one side of the global mean:

$$m_1 = \frac{\int_{-\infty}^{\bar{x}} f(x')x'\,\mathrm{d}x'}{\int_{-\infty}^{\bar{x}} f(x')\,\mathrm{d}x'} \qquad m_2 = \frac{\int_{\bar{x}}^{+\infty} f(x')x'\,\mathrm{d}x'}{\int_{\bar{x}}^{+\infty} f(x')\,\mathrm{d}x'}. \tag{C.5}$$

The denominators normalize the truncated pdfs. (Note that, since we have replaced the samples by their asymptotic pdf, the expressions in equation (C.5) represent the estimate means.) Obviously, $m = m_2 - m_1$.

Combining equations (C.1) and (C.5) results in

$$m_1 = \frac{\frac{1}{2}\left[\int_{-\infty}^{\bar{x}} F(x'-\mu_1)x'\,\mathrm{d}x' + \int_{-\infty}^{\bar{x}} F(x'-\mu_2)x'\,\mathrm{d}x'\right]}{\frac{1}{2}\left[\int_{-\infty}^{\bar{x}} F(x'-\mu_1)\,\mathrm{d}x' + \int_{-\infty}^{\bar{x}} F(x'-\mu_2)\,\mathrm{d}x'\right]} = \frac{A_1 + B_1}{C_1 + D_1} \tag{C.6}$$

$$m_2 = \frac{\frac{1}{2}\left[\int_{\bar{x}}^{+\infty} F(x'-\mu_1)x'\,\mathrm{d}x' + \int_{\bar{x}}^{+\infty} F(x'-\mu_2)x'\,\mathrm{d}x'\right]}{\frac{1}{2}\left[\int_{\bar{x}}^{+\infty} F(x'-\mu_1)\,\mathrm{d}x' + \int_{\bar{x}}^{+\infty} F(x'-\mu_2)\,\mathrm{d}x'\right]} = \frac{A_2 + B_2}{C_2 + D_2}. \tag{C.7}$$

Consider A_2 in equation (C.7). A change of variable $u' = x' - \mu_1$ results in

$$A_2 = \int_{\frac{\delta}{2}}^{+\infty} F(u')u'\,\mathrm{d}u' + \mu_1 \int_{\frac{\delta}{2}}^{+\infty} F(u')\,\mathrm{d}u'. \tag{C.8}$$

(Remember that $\delta = \mu_2 - \mu_1$.) Since $\int_{\frac{\delta}{2}}^{+\infty} = \int_0^{\infty} - \int_0^{\frac{\delta}{2}}$, we can use the assumption that δ is small to approximate the above integrals

$$\int_0^{\frac{\delta}{2}} F(u')\,\mathrm{d}u' \approx F(0)\int_0^{\frac{\delta}{2}} \mathrm{d}u' = \frac{\delta}{2}F(0) \tag{C.9}$$

$$\int_0^{\frac{\delta}{2}} F(u')u'\,\mathrm{d}u' \approx F(0)\int_0^{\frac{\delta}{2}} u'\,\mathrm{d}u' = \frac{\delta^2}{8}F(0). \tag{C.10}$$

Equation (C.8) now reads

$$A_2 = \int_0^{+\infty} F(u')u'\,\mathrm{d}u' - \frac{\delta^2}{8}F(0) + \mu_1\left[\int_0^{+\infty} F(u')\,\mathrm{d}u' - \frac{\delta}{2}F(0)\right]. \tag{C.11}$$

In a completely similar way, we can find the following:

$$B_2 = \int_0^{+\infty} F(u')u'\,\mathrm{d}u' - \frac{\delta^2}{8}F(0) + \mu_2\left[\int_0^{+\infty} F(u')\,\mathrm{d}u' + \frac{\delta}{2}F(0)\right] \tag{C.12}$$

$$C_2 = \int_0^{+\infty} F(u')\,\mathrm{d}u' - \frac{\delta}{2}F(0) \tag{C.13}$$

$$D_2 = \int_0^{+\infty} F(u')\,\mathrm{d}u' + \frac{\delta}{2}F^2(0). \tag{C.14}$$

For m_1, we will have

$$A_1 = \int_{-\infty}^{0} F(u')u'\,\mathrm{d}u' + \frac{\delta^2}{8}F(0) + \mu_1\left[\int_{-\infty}^{0} F(u')\,\mathrm{d}u' + \frac{\delta}{2}F(0)\right] \tag{C.15}$$

$$B_1 = \int_{-\infty}^{0} F(u')u'\,\mathrm{d}u' + \frac{\delta^2}{8}F(0) + \mu_2\left[\int_{-\infty}^{0} F(u')\,\mathrm{d}u' - \frac{\delta}{2}F(0)\right] \tag{C.16}$$

$$C_1 = \int_{-\infty}^{0} F(u')\,\mathrm{d}u' + \frac{\delta}{2}F(0) \tag{C.17}$$

$$D_1 = \int_{-\infty}^{0} F(u')\,\mathrm{d}u' - \frac{\delta}{2}F(0). \tag{C.18}$$

Putting everything together we find

$$m = m_2 - m_1 = \frac{(A_2 + B_2)(C_1 + D_1) - (A_1 + B_1)(C_2 + D_2)}{(C_1 + D_1)(C_2 + D_2)} \tag{C.19}$$

$$= \frac{\int_0^{+\infty} F(u')u' \,\mathrm{d}u' + \frac{F(0)}{8}\delta^2}{\int_0^{+\infty} F(u') \,\mathrm{d}u' \int_{-\infty}^{0} F(u') \,\mathrm{d}u'}. \tag{C.20}$$

Thus, there is no linear term in the expression for $\hat{m}$.

C.2. Second-moment estimator method

Since m is, by definition, an expectation value, it can be found by increasing N to infinity, where the variance σ_R approaches zero. At this limit, measurement moments can be replaced by moments of pdfs they are sampled from:

$$m^2 = V_2(N \to \infty) = \int_{-\infty}^{+\infty} f(u')u'^2 \,\mathrm{d}u' - \left[\int_{-\infty}^{+\infty} f(u')u' \,\mathrm{d}u'\right]^2. \tag{C.21}$$

with $f(\mathrm{x})$ defined in (C.1). It is easy to see that

$$\int_{-\infty}^{+\infty} f(u')u' \,\mathrm{d}u' = \frac{\mu_1 + \mu_2}{2} \tag{C.22}$$

$$\int_{-\infty}^{+\infty} f(u')u'^2 \,\mathrm{d}u' = \frac{\mu_1^2 + \mu_2^2}{2} + \int_{-\infty}^{+\infty} F(u')u'^2 \,\mathrm{d}u'. \tag{C.23}$$

As a result,

$$m^2 = \int_{-\infty}^{+\infty} F(u')u'^2 \,\mathrm{d}u' + \frac{\delta^2}{4}. \tag{C.24}$$

Clearly, for small δ

$$m \simeq \left[\int_{-\infty}^{+\infty} F(u')u'^2 \,\mathrm{d}u'\right]^{1/2} + \frac{\delta^2}{8\left[\int_{-\infty}^{+\infty} F(u')u'^2 \,\mathrm{d}u'\right]^{1/2}} \tag{C.25}$$

which contains no linear term.

References

Akaike H 1974 *IEEE Trans. Autom. Control* **19** 716
Bozdogan H and Sclove S L 1984 *Ann. Inst. Stat. Math.* **36** 163
Chen J 1994 *Can. J. Stat.* **22** 387
Chen J and Cheng P 1997 *Can. J. Stat.* **25** 389
Dayan P and Abbott L F 2001 *Theoretical Neuroscience* (Cambridge, MA: MIT Press)
Dempster A, Laird N and Rubin D 1977 *J. R. Stat. Soc.* B **39** 1
Duda R O, Hart P E and Stork D G 2000 *Pattern Classification* (New York: Wiley)
Everitt B S and Hand D J 1981 *Finite Mixture Distributions* (New York: Chapman and Hall)
Frayley C and Raftery A 1998 *Comput. J.* **41** 578
Furman W D and Lindsay B G 1994 *Comput. Stat. Data Anal.* **17** 473
Goffinet B, Loisel P and Laurent B 1992 *Biometrika* **79** 842
Gordon A D 1999 *Classification* (New York/Boca Raton FL: Chapman and Hall/CRC)
Ghosh J H and Sen P K 1985 *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* vol 2 p 789
Hardy A 1996 *Comput. Stat. Data Anal.* **23** 83
Hartigan J A 1985a *J. Classif.* **2** 63

Hartigan J A 1985b *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* vol 2 p 807
Lloyd S P 1982 *IEEE Trans. Inform. Theory* **28** 129
Lo Y, Mendell N R and Rubin D B 2001 *Biometrika* **88** 767
Mangin B, Goffinet B and Elsen J M 1993 *Biometrical J.* **35** 771
McLachlan G J and Krishnan T 1997 *The EM Algorithm and Extensions* (New York: Wiley)
McLachlan G J and Peel D 2000 *Finite Mixture Models* (New York: Wiley)
Milligan G W and Cooper M C 1985 *Psychometrika* **50** 159
Pearson K 1894 *Phil. Trans. R. Soc.* A **185** 71
Polymenis A and Titterington D M 1998 *Stat. Prob. Lett.* **38** 295
Schwarz G 1978 *Ann. Stat.* **6** 461
Sclove S L 1987 *Psychometrika* **52** 333
Tibshirani R, Walther G and Hastie T 2001 *J. R. Stat. Soc.* B **63** 411
Titterington D M, Smith A F M and Makov U E 1985 *Statistical Analysis of Finite Mixture Distributions* (New York: Wiley)