Notes on Big MRA

William Leeb

1 Setup

We have the following model. x is a signal of length L. For large n, we observe the vector $Y \in \mathbb{R}^n$, which is of the form

$$Y = G * x + N \tag{1}$$

where G is a sum of diracs at locations at least $2L$ apart, and N is a vector of n iid Gaussians. The goal is to recover x.

We will assume that the first and last copies of x are at least L places removed from the boundaries of the interval $[1, n + L]$ (if not we can just zero-pad the interval). We will also denote by $I_1, \ldots, I_J$ the J subintervals containing the signal; and take $I_j = [a_j, b_j]$. When we take the limit $n \to \infty$, we'll let $J = J_n$ grow with n, as is natural. We will see that we require $J_n = \Omega(n)$ in order for the limits to not vanish (and obviously $J_n = O(n)$ too).

For the moments computation, I'm not just computing the expected value. I'm actually showing convergence almost surely as $n \to \infty$. I probably made a mistake in the final formulas since I haven't check them carefully yet; the main point I want to record right now is how to get the bookkeeping correct for dealing with the noise terms, namely, we need to break the averages into averages of averages, each with iid terms.

2 First moment

We have:

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n/L}\sum_{j=1}^{J_n}\frac{1}{L}\sum_{i=1}^{L} x_i + \frac{1}{n}\sum_{i=1}^{n} N_i \to \gamma \cdot \bar{x}, \tag{2}$$

where the limit is almost surely (we've used the strong law of large numbers), where $\gamma = \lim_{n\to\infty} J_n L/n$ is the fraction of the observations containing the signal.

3 Second moment

We fix a value Δ between 0 and $L-1$. We will compute the second moment:

$$M_2(\Delta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-\Delta} Y_i Y_{i+\Delta}. \tag{3}$$

3.1 Clean signal without noise

First, if there is no noise, and under the wide spacing assumption, we can break the sum into J_n different sums, one for each copy of x embedded in the signal:

$$\begin{aligned} M_2(\Delta) &= \frac{1}{n} \sum_{j=1}^{J_n} \sum_{i=a_j}^{b_j-\Delta} Y_i Y_{i+\Delta} = \frac{1}{n} \sum_{j=1}^{J_n} \sum_{i=1}^{L-\Delta} x_i x_{i+\Delta} \\ &= \frac{J_n L}{n} \frac{1}{L} \sum_{i=1}^{L-\Delta} x_i x_{i+\Delta} \to \gamma \cdot R_2(\Delta), \end{aligned} \tag{4}$$

where here we have defined $R_2(\Delta)$ as element Δ of the autocorrelation of x.

3.2 Pure noise without signal

Here, $Y_i = N_i$. We first fix Δ between 1 and $L-1$. Break up the sum into Δ terms as follows:

$$\begin{aligned} M_2(\Delta) &= \frac{1}{n} \sum_{i=1}^{n} N_i N_{i+\Delta} \\ &= \frac{1}{\Delta} \sum_{m=0}^{\Delta-1} \frac{1}{n/\Delta} \sum_{j=1}^{n/\Delta} N_{j+(j-1)\Delta+m} N_{j+j\Delta+m}. \end{aligned} \tag{5}$$

Each term $\frac{1}{n/\Delta} \sum_{j=1}^{n/\Delta} N_{j\Delta+m} N_{(j+1)\Delta+m}$ is an average of n/Δ iid terms with expectation zero, and so converges to 0 a.s. as $n \to \infty$. (Of course, it's not exactly n/Δ terms; there will be finitely many terms unaccounted for this way, but these are negligible).

If $\Delta = 0$, then the computation is even easier:

$$M_2(\Delta) = \frac{1}{n} \sum_{i=1}^{n} N_i^2 \xrightarrow{a.s.} \sigma^2. \tag{6}$$

So $M_2(\Delta) = \sigma^2$ if $\Delta = 0$, and 0 otherwise.

3.3 Signal plus noise

We will denote by $\mathcal{X} = x * G$, so $Y = \mathcal{X} + N$. Then the second moment of the signal plus noise is:

$$M_2(\Delta, Y) = M_2(\Delta, \mathcal{X}) + M_2(\Delta, N) + \frac{1}{n}\sum_{i=1}^{n-\Delta} \mathcal{X}_i N_{i+\Delta} + \frac{1}{n}\sum_{i=1}^{n-\Delta} \mathcal{X}_{i+\Delta} N_i. \quad (7)$$

The law of large numbers says the cross terms vanish as $n \to \infty$. So the limit is simply:

$$M_2(\Delta, Y) = \begin{cases} \gamma \cdot R_2(\Delta), & \text{if } \Delta > 0; \\ \gamma \cdot R_2(\Delta) + \sigma^2, & \text{if } \Delta = 0. \end{cases} \quad (8)$$

4 Third moments

The same idea lets us compute the third moments. We fix two indices $\Delta_1 < \Delta_2$, and define:

$$M_3(\Delta_1, \Delta_2) = \sum_{i=1}^{n-\Delta_2} Y_i Y_{i+\Delta_1} Y_{i+\Delta_2}. \quad (9)$$

4.1 Pure noise, no signal

Again, the idea is to break up the big average over n terms into a sum of Δ_2 averages, each of $\approx n/\Delta_2$ independent terms. We write:

$$\begin{aligned} &M_3(\Delta_1, \Delta_2) \\ &= \frac{1}{\Delta_2}\sum_{m=0}^{\Delta_2 - 1} \frac{1}{(n/\Delta_2)} \sum_{j=1}^{n/\Delta_2} N_{j+(j-1)\Delta_2+m} N_{j+\Delta_1+(j-1)\Delta_2+m} N_{j+j\Delta_2+m}. \end{aligned} \quad (10)$$

Each of the Δ_2 terms indexed by m converges is an average of n/Δ_2 independent terms with mean zero, and so converges a.s. to zero as $n \to \infty$.

4.2 Clean signal, no noise

Again, this is just like the second moment case. Write:

$$\begin{aligned} M_3(\Delta_1, \Delta_2) &= \frac{1}{n}\sum_{j=1}^{J_n} \sum_{i=a_j}^{b_j - \Delta_2} Y_i Y_{i+\Delta_1} Y_{i+\Delta_2} = \frac{1}{n}\sum_{j=1}^{J_n} \sum_{i=1}^{L-\Delta_2} x_i x_{i+\Delta_1} x_{i+\Delta_2} \\ &= \frac{J_n L}{n}\frac{1}{L}\sum_{i=1}^{L-\Delta} x_i x_{i+\Delta_1} x_{i+\Delta_2} \to \gamma \cdot R_3(\Delta_1, \Delta_2), \end{aligned} \quad (11)$$

where here we have defined $R_3(\Delta_1, \Delta_2)$ as element (Δ_1, Δ_2) of the third moment of x.

4.3 Signal plus noise

We have:

$$M_3(\Delta_1, \Delta_2, Y) = M_3(\Delta_1, \Delta_2, \mathcal{X}) + M_3(\Delta_1, \Delta_2, N) + \text{cross terms}. \tag{12}$$

If $0 < \Delta_1 < \Delta_2$, then all of the cross terms can be shown to go to zero by the same kind of argument we've used several times before.

If $\Delta_1 = \Delta_2 = \Delta > 0$, there is a surviving cross-term, namely:

$$\frac{1}{n}\sum_{i=1}^{n-\Delta} \mathcal{X}_i N_{i+\Delta}^2 = \frac{1}{(n/L)}\sum_{j=1}^{J_n} \frac{1}{L}\sum_{i=a_j}^{b_j} \mathcal{X}_i N_{i+\Delta}^2. \tag{13}$$

Each of $S_j \equiv \frac{1}{L}\sum_{i=a_j}^{b_j} \mathcal{X}_i N_{i+\Delta}^2 \sim \frac{1}{L}\sum_{i=1}^{L} x_i \varepsilon_i^2$, where $\varepsilon_i \sim N(0, \sigma^2)$; and they are independent random variables with mean $\mathbb{E}[S_j] = \sigma^2 \cdot \bar{x}$. So by the law of large numbers,

$$\frac{1}{(n/L)}\sum_{j=1}^{J_n} \frac{1}{L}\sum_{i=a_j}^{b_j} \mathcal{X}_i N_{i+\Delta}^2 \to \gamma \cdot \sigma^2 \cdot \bar{x}. \tag{14}$$

Similarly, if $\Delta_2 > \Delta_1 = 0$, then there is a surviving cross-term which converges to $\gamma \cdot \sigma^2 \cdot \bar{x}$ as well. Finally, if $\Delta_1 = \Delta_2 = 0$, then the cross-term converges to $3 \cdot \gamma \cdot \sigma^2 \cdot \bar{x}$. So in summary, for all $0 \le \Delta_1 \le \Delta_2 \le L-1$, we have:

$$M_3(\Delta_1, \Delta_2, Y) = \begin{cases} \gamma \cdot R_3(\Delta_1, \Delta_2), & \text{if } 0 < \Delta_1 < \Delta_2; \\ \gamma \cdot R_3(\Delta_1, \Delta_2) + \gamma \cdot \sigma^2 \cdot \bar{x}, & \text{if } \Delta_2 > \Delta_1 = 0 \text{ or } \Delta_1 = \Delta_2 > 0; \\ \gamma \cdot R_3(0,0) + 3 \cdot \gamma \cdot \sigma^2 \cdot \bar{x} & \text{if } \Delta_1 = \Delta_2 = 0. \end{cases} \tag{15}$$

5 Attempt at EM

Let's derive the EM algorithm, where we treat the locations a_j as random latent variables. A simple model is that the a_j are chosen independently and uniformly at random (though take note that this permits overlaps of the signal copies).

We will denote by $G = \sum_{j=1}^{J} \delta_{a_j}$ the random vector of signal locations. We'll assume the number of signals J is known. The generative model is:

$$p(Y, G|x) = \frac{\exp\{-\|Y - G * x\|^2/(2\sigma^2)\}}{(2\pi\sigma^2)^{n/2}} \cdot \binom{n}{J}^{-1}. \tag{16}$$

The log-likelihood is then (up to an additive constant):

$$\mathcal{L}(x; Y, G) \propto -\|Y - G * x\|^2. \tag{17}$$

The probability distribution of G, given Y and x, is:

$$p(G|Y, x) = \frac{p(Y, G|x)}{p(Y|x)} \propto \exp\{-\|Y - G * x\|^2/(2\sigma^2)\}. \tag{18}$$

If $x^{(t)}$ is a guess for x on the t^{th} iterate of EM, then we define the weights:

$$w_t(G) = p(G|Y, x^{(t)}) = \exp\{-\|Y - G * x^{(t)}\|^2/(2\sigma^2)\}, \tag{19}$$

and the Q-function:

$$\begin{aligned} Q(x|x^{(t)}) &= \mathbb{E}_{G|Y,x^{(t)}}\left[\mathcal{L}(x^{(t)}; Y, G)\right] \\ &\propto -\sum_G \|Y - G * x\|^2 \cdot w_t(G). \end{aligned} \tag{20}$$

The EM algorithm defines the next guess of x to be:

$$x^{(t+1)} = \arg\max_x Q(x|x^{(t)}) = \arg\min_x \sum_G \|Y - G * x\|^2 \cdot w_t(G). \tag{21}$$

But finding this minimum is hard, since the sum is over all $\binom{n}{J}$ possible values of G; if $J \sim n$, as we expect is necessary to have any chance of recovery (this was suggested by the moments method), the number of terms will grow super-exponentially with n.

6 Estimating γ, using σ and the moments

Using the moments calculation, we will derive an equation relating γ and σ. If σ were known (and $\mathbf{1}^\top x \neq 0$), this permits us to then estimate the unscaled moments of x, which in turn lets us fit x to its moments.

It will be notationally simpler to instead estimate $\beta \equiv \gamma \cdot L = J/n$. Also, we'll define the unnormalized moments of x by $T_i = L \cdot R_i$, for $i = 2, 3$. From the first moment we estimate $\beta \cdot (\mathbf{1}^\top x)$, and hence also its square:

$$\begin{aligned} A \equiv (\beta \cdot \mathbf{1}^\top x)^2 &= \beta^2 \left(\sum_{i=1}^{L} x_i^2 + 2 \sum_{i=1}^{L} \sum_{j=i+1}^{L} x_i x_j \right) \\ &= \beta^2 \left(T_2(0) + 2 \sum_{i=1}^{L} \sum_{\Delta=1}^{L-i} x_i x_{i+\Delta} \right) \\ &= \beta^2 \left(T_2(0) + 2 \sum_{i=1}^{L} \sum_{\Delta=1}^{L-1} x_i x_{i+\Delta} \mathbf{1}_{(\Delta \leq L-i)} \right) \\ &= \beta^2 \left(T_2(0) + 2 \sum_{\Delta=1}^{L-1} \sum_{i=1}^{L-\Delta} x_i x_{i+\Delta} \right) \\ &= \beta^2 \left(T_2(0) + 2 \sum_{\Delta=1}^{L-1} T_2(\Delta) \right). \end{aligned} \tag{22}$$

On the other hand, from the second moment we can estimate $\beta \cdot T_2(\Delta)$ for all $1 \leq \Delta \leq L-1$, and $\beta \cdot T_2(0) + \sigma^2$, and consequently $B \equiv \beta \cdot (T_2(0) +$

$2\sum_{\Delta=1}^{L} T_2(\Delta)) + \sigma^2$. Taking the ratio, we obtain the equation:

$$\frac{1}{\beta} + \frac{\sigma^2}{A} = \frac{B}{A}. \tag{23}$$

If we knew σ, we can solve for β:

$$\beta = \frac{A}{B - \sigma^2}. \tag{24}$$

(Of course, this only works if $B \neq \sigma^2$, or equivalently if $\mathbf{1}^\top x \neq 0$.)

7 A pair of quadratics satisfied by β

It turns out that β satisfies two independent quadratic equations, defined in terms of the observed third-order moments. The coefficients of these quadratics depend only on the observed moments, not on σ; so we can estimate β (and hence γ) consistently, even without knowing the noise level. We can then estimate σ, and from them, recover the moments of x, which in turn lets us solve for x.

Of course, for finite n and large σ the formulas for β and σ will be noisy. Consequently, I suggest instead adding them as variables to the non-linear least squares problem; that is, solve for all of x, β and $\nu \equiv \sigma^2$ by fitting their moments to the observed moments via least-squares.

7.1 The first quadratic equation

The previous section gives the equation:

$$\beta = \frac{A}{B - \sigma^2}. \tag{25}$$

Now we move up to third-order moments. From the first-order moments, we can estimate $\beta \cdot \mathbf{1}^\top x$; and from the second-order moments we can estimate $\beta \cdot \|x\|^2 + \sigma^2$. Taking the product, we have:

$$\begin{aligned} C &\equiv (\beta \cdot \mathbf{1}^\top x) \cdot (\beta \cdot \|x\|^2 + \sigma^2) \\ &= \beta^2 \left(T_3(0,0) + \sum_{\Delta=1}^{L-1} T_3(0,\Delta) + \sum_{\Delta=1}^{L-1} T_3(\Delta,\Delta) \right) + \sigma^2 \cdot (\beta \cdot \mathbf{1}^\top x) \\ &= \beta^2 \cdot \omega + \sigma^2 \cdot E, \end{aligned} \tag{26}$$

where $\omega = T_3(0,0) + \sum_{\Delta=1}^{L-1} T_3(0,\Delta) + \sum_{\Delta=1}^{L-1} T_3(\Delta,\Delta)$, and

$$E = \beta \cdot \mathbf{1}^\top x. \tag{27}$$

From the observed third-order terms themselves, we can estimate:

$$D \equiv \beta \cdot \omega + (2L+1) \cdot \sigma^2 \cdot (\beta \cdot \mathbf{1}^\top x) = \beta \cdot \omega + \sigma^2 \cdot F, \tag{28}$$

where

$$F = (2L+1) \cdot (\beta \cdot \mathbf{1}^\top x). \tag{29}$$

Consequently:

$$D \cdot \beta - \sigma^2 \cdot \beta \cdot F = \beta^2 \cdot w = C - \sigma^2 \cdot E. \tag{30}$$

Using $\sigma^2 = B - A/\beta$ and rearranging, we get:

$$(D - BF) \cdot \beta^2 + (AF + BE - C) \cdot \beta - AE \equiv \mathcal{A}\beta^2 + \mathcal{B}\beta + \mathcal{C} = 0. \tag{31}$$

We can simplify the expressions a little bit. First, we observe that:

$$\mathcal{C} = -AE = -(\beta \cdot \mathbf{1}^\top x)^3 = -E^3 \equiv -G. \tag{32}$$

Second, we can simplify $\mathcal{B}$:

$$\mathcal{B} = AF + BE - C = (2L+1) \cdot G + 2E\beta \sum_{\Delta=1}^{L-1} T_2(\Delta). \tag{33}$$

7.2 The second quadratic equation

It is a straightforward computation to show that:

$$\begin{aligned} \phi &\equiv \left(\sum_{i=1}^{L} x_i \right)^3 \\ &= T_3(0,0) + 3 \sum_{\Delta=1}^{L} T_3(\Delta,\Delta) + 3 \sum_{\Delta=1}^{L} T_3(0,\Delta) + 6 \sum_{1 \le \Delta_1 < \Delta_1 \le L-1} T_3(\Delta_1,\Delta_2). \end{aligned} \tag{34}$$

The quantity $E = \beta \cdot \mathbf{1}^\top x$ is observed, and hence so is $G \equiv E^3 = \beta^3 \cdot \phi$. On the other hand, from the observed third moments we can directly estimate:

$$\begin{aligned} H &\equiv M_3(0,0) + 3 \sum_{\Delta=1}^{L} M_3(\Delta,\Delta) + 3 \sum_{\Delta=1}^{L} M_3(0,\Delta) + 6 \sum_{1 \le \Delta_1 < \Delta_1 \le L-1} M_3(\Delta_1,\Delta_2) \\ &= \beta \cdot \phi + (6L-3) \cdot E \cdot \sigma^2 = \beta \cdot \phi + O \cdot \sigma^2, \end{aligned} \tag{35}$$

where the quantity

$$O = (6L-3) \cdot E = (6L-3) \cdot \beta \cdot \mathbf{1}^\top x \tag{36}$$

is also observed. Taking the ratio of G and H, we get:

$$\frac{H}{G} = \frac{\beta \cdot \phi + O \cdot \sigma^2}{G} = \frac{\beta \cdot \phi}{\beta^3 \cdot \phi} + \frac{O \cdot \sigma^2}{G} = \frac{1}{\beta^2} + \frac{O}{G} \cdot \sigma^2, \tag{37}$$

and solving for σ^2 in terms of β, we get:

$$\sigma^2 = \frac{H}{O} - \frac{G}{O} \cdot \frac{1}{\beta^2}. \tag{38}$$

On the other hand, we already know:

$$\sigma^2 = B - \frac{A}{\beta}. \tag{39}$$

Combining these two equations, we arrive at the quadratic:

$$(B - H/O) \cdot \beta^2 - A \cdot \beta + (G/O) \equiv \mathcal{D}\beta^2 + \mathcal{E}\beta + \mathcal{F} = 0. \tag{40}$$

7.3 Independence of the quadratics

We can show that for generic x, the two quadratics are independent; and in fact, for all sufficiently large β, they are independent for any x (not just generic). By "generic" I mean on a set with complement of Lebesgue measure 0.

To see this, it's enough to show that the ratio of coefficients is not the same. We have:

$$\frac{\mathcal{B}}{\mathcal{C}} = -(2L+1) - \frac{2\beta \sum_{\Delta \geq 1} T_2(\Delta)}{E^2} \tag{41}$$

and

$$\frac{\mathcal{E}}{\mathcal{F}} = -(6L-3). \tag{42}$$

Suppose the quadratics are dependent. Then $\mathcal{B}/\mathcal{C} = \mathcal{E}/\mathcal{F}$, or

$$2L + 1 + \frac{2\beta \sum_{\Delta \geq 1} T_2(\Delta)}{E^2} = 6L - 3 \tag{43}$$

or equivalently

$$\begin{aligned} 4(L-1)\beta^2(\mathbf{1}^\top x)^2 &= 4(L-1)E^2 \\ &= 2\beta \sum_{\Delta \geq 1} T_2(\Delta) \\ &= \beta((\mathbf{1}^\top x)^2 - \|x\|^2). \end{aligned} \tag{44}$$

Rearranging we get:

$$(4(L-1)\beta - 1)(\mathbf{1}^\top x)^2 + \|x\|^2 = 0. \tag{45}$$

This describes a quadratic which has measure zero. But we can actually say something more. Since $(\mathbf{1}^\top x)^2 \leq L\|x\|^2$, we must have:

$$1 - 4(L-1)\beta = \frac{\|x\|^2}{(\mathbf{1}^\top x)^2} \geq \frac{1}{L} \tag{46}$$

i.e.

$$(L-1)(1 - 4L\beta) = L - 1 - 4(L-1)L\beta \geq 0 \tag{47}$$

so that we need $\beta \leq 1/(4L)$. In other words, if $\beta \geq 1/(4L)$, then the quadratics are independent.

8 Relation to system identification

Our problem can be interpreted as a special case of the problem of "system identification" in the field of signal processing. In this class of problems, we observe a length n signal Y given by:

$$Y = x * W + \varepsilon, \tag{48}$$

where x is an unknown linear filter of length L; W is an unknown random "input" sequence; and ε is random additive noise. The problem has also been studied in the case of a known input W [9, 5, 3]. The goal of this problem is to estimate the "system" vector x. The question of identifiability of x under this observation model is addressed for certain non-Gaussian W in [2], and for certain Gaussian inputs W in [7]. In the special case where W is a random, unknown sequence of spikes satisfying the spacing requirement we obtain our model in the $K = 1$ case of a single vector.

Likelihood-based methods seek to maximize the likelihood function for x, given the observed signal Y. Solving this optimization exactly is typically not tractable, and so heuristic methods are used instead. One proposed heuristic for computing the maximization is to use Markov Chain Monte Carlo; in special cases, including the case where W is binary, expectation-maximization has been used [4]. The EM method is based upon a certain "forward-backward" procedure used in hidden Markov models, described in [10].

Other problems solve for parameterized models of x; for instance, see [1]. That same paper also considers multiple distinct signals x embedded in the same observation vector Y, as in our heterogeneity framework. Their proposed solution is a Markov Chain Monte Carlo algorithm designed for their specific parametrized problem.

Because likelihood methods are computationally expensive, methods based on recovery from moments and cumulants, which are akin to our method, have also been previously used for system identification. The question of identifiablity of x from third-order statistics with non-Gaussian W was considered in [8]. A method using fourth-order cumulants is described in [11], while a method using third-order cumulants is described in [6].

9 Impossibility of detection

Let's consider an even easier problem, in which someone hands us intervals of length L that either contain the full signal plus noise, or contain just noise, and we are asked to determine which intervals contain signal. Let's assume too that we know the signal x, that we know the probability q that one of the intervals contains a signal, and that we know the noise variance σ. Since we have the full generative model, we can always generate more data as needed; therefore it's enough to consider the hardness of deciding whether a single interval contains signal or not.

Let's abstract this problem. We have two known vectors θ_0 and θ_1 in $\mathbb{R}^L$. There is a random variable ε taking values 0 or 1 with probabilities q and $1-q$, respectively. We observe a random vector $X \in \mathbb{R}^L$ with the following distribution: $X|(\varepsilon = i) \sim N(\theta_i, \sigma^2)$; in other words, if $\varepsilon = 0$, then $X \sim N(\theta_0, \sigma^2)$, and if $\varepsilon = 1$, then $X \sim N(\theta_1, \sigma^2)$.

We observe X, and our task is to predict ε. How well can we do it? First, let's assume without loss of generality that $q \geq 1-q$. Then if we make the constant prediction $\hat{\varepsilon} \equiv 0$, we will be correct with probability q. So the question is, can we do better than this? We will prove that as $\sigma \to \infty$, the answer is no. More precisely:

Proposition 9.1. *Let $\hat{\varepsilon}$ be any predictor of ε. Then*

$$\lim_{\sigma \to \infty} \mathbb{P}[\hat{\varepsilon} = \varepsilon] \leq q. \tag{49}$$

To begin the proof, we use a variant of the Neyman-Pearson Lemma to derive the best (deterministic) predictor $\hat{\varepsilon}$. We take any predictor $\hat{\varepsilon}$, and we let S be the set of X's where $\hat{\varepsilon} = 1$. Then the probability that $\hat{\varepsilon}$ fails is:

$$\begin{aligned}
\mathbb{P}[\hat{\varepsilon} \neq \varepsilon] &= q\mathbb{P}_0[\hat{\varepsilon} = 1] + (1-q)\mathbb{P}_1[\hat{\varepsilon} = 0] \\
&= q\mathbb{P}_0[\hat{\varepsilon} = 1] + (1-q)(1 - \mathbb{P}_1[\hat{\varepsilon} = 1]) \\
&= q\mathbb{P}_0[\hat{\varepsilon} = 1] + (1-q) - (1-q)\mathbb{P}_1[\hat{\varepsilon} = 1] \\
&= (1-q) + \int_S (qf_0(x) - (1-q)f_1(x))dx,
\end{aligned} \tag{50}$$

where $f_i(x)$ is the normal density with mean θ_i and variance σ^2, and the notation $\mathbb{P}_i$ means probability conditional on the event $\varepsilon = \theta_i$; that is, $\mathbb{P}_i[A] = \mathbb{P}[A|\varepsilon = i]$. The integral (50) is minimized by taking S to be the set:

$$S = \left\{ x : \frac{f_0(x)}{f_1(x)} \leq \frac{1-q}{q} \right\}. \tag{51}$$

In other words, if $\Lambda(x) = f_0(x)/f_1(x)$ and $b = (1-q)/q$, then the best predictor of ε based on X is:

$$\varepsilon = \begin{cases} 1, & \text{if } \Lambda(x) \leq b \\ 0, & \text{if } \Lambda(x) \geq b \end{cases}. \tag{52}$$

Taking logarithms, the set S can be rewritten as the set of x where:

$$-\|x - \theta_0\|^2 \leq -\|x - \theta_1\|^2 + 2\sigma^2 \log(b), \tag{53}$$

or equivalently

$$\langle x, \theta_1 - \theta_0 \rangle \geq \frac{\|\theta_1\|^2 - \|\theta_0\|^2}{2} - \sigma^2 \log(b). \tag{54}$$

Now let's compute the probability of failure conditional on the event $\varepsilon = 0$. In this case, failure occurs when $X \in S$. Since $X|(\varepsilon = 0) \sim N(\theta_0, \sigma^2)$, we can write $X|(\varepsilon = 0) = \sigma Z + \theta_0$, where $Z \sim N(0,1)$. Consequently,

$$\begin{aligned}\langle X, \theta_1 - \theta_0 \rangle &= \sigma \langle Z, \theta_1 - \theta_0 \rangle + \langle \theta_0, \theta_1 - \theta_0 \rangle \\ &= \sigma \langle Z, \theta_1 - \theta_0 \rangle + \langle \theta_0, \theta_1 \rangle - \|\theta_0\|^2 \end{aligned} \tag{55}$$

and failure occurs when

$$\begin{aligned}\sigma \langle Z, \theta_1 - \theta_0 \rangle &\geq \frac{\|\theta_1\|^2 + \|\theta_0\|^2}{2} - \langle \theta_0, \theta_1 \rangle - \sigma^2 \log(b) \\ &= \frac{1}{2}\|\theta_1 - \theta_0\|^2 - \sigma^2 \log(b) \end{aligned} \tag{56}$$

or equivalently

$$Y \geq \frac{c}{\sigma} - \sigma \log(b) \tag{57}$$

where $c = \|\theta_1 - \theta_0\|^2/2$ and $Y = \langle Z, \theta_1 - \theta_0 \rangle \sim N(0, \|\theta_1 - \theta_0\|^2)$. Just for simplicity let's assume $\|\theta_1 - \theta_0\| = 1$, so that $Y \sim N(0,1)$. So:

$$\mathbb{P}_0[\hat{\varepsilon} = 1] = \mathbb{P}\left[Y \geq \frac{c}{\sigma} - \sigma^2 \log(b)\right], \quad Y \sim N(0,1). \tag{58}$$

Similarly,

$$\mathbb{P}_1[\hat{\varepsilon} = 0] = \mathbb{P}\left[Y \geq \frac{c}{\sigma} + \sigma^2 \log(b)\right], \quad Y \sim N(0,1). \tag{59}$$

So the overall probability of failure is:

$$\mathbb{P}[\hat{\varepsilon} \neq \varepsilon] = q\mathbb{P}\left[Y \geq \frac{c}{\sigma} - \sigma^2 \log(b)\right] + (1-q)\mathbb{P}\left[Y \geq \frac{c}{\sigma} + \sigma^2 \log(b)\right]. \tag{60}$$

Now, if $q = 1/2$, the $b = 1$ and $\log(b) = 0$. Then the probability of failure is simply:

$$\mathbb{P}\left[Y \geq \frac{c}{\sigma}\right] \longrightarrow \frac{1}{2} = q \quad \text{as} \quad \sigma \to \infty. \tag{61}$$

On the other hand, if $q > 1 - q$, then $\log(b) < 0$. Then

$$\mathbb{P}\left[Y \geq \frac{c}{\sigma} - \sigma^2 \log(b)\right] \longrightarrow 0 \tag{62}$$

while

$$\mathbb{P}\left[Y \geq \frac{c}{\sigma} + \sigma^2 \log(b)\right] \longrightarrow 1 \tag{63}$$

as $\sigma \to \infty$. Consequently,

$$\begin{aligned}\mathbb{P}[\hat{\varepsilon} \neq \varepsilon] &= q\mathbb{P}\left[Y \geq \frac{c}{\sigma} - \sigma^2 \log(b)\right] + (1-q)\mathbb{P}\left[Y \geq \frac{c}{\sigma} + \sigma^2 \log(b)\right] \\ &\longrightarrow 1 - q \quad \text{as} \quad \sigma \to \infty. \end{aligned} \tag{64}$$

That is, the probability of success converges to q, which completes the proof.

References

[1] Christophe Andrieu, Éric Barat, and Arnaud Doucet. Bayesian deconvolution of noisy filtered point processes. *IEEE Transactions on Signal Processing*, 49(1):134–146, 2001.

[2] Albert Benveniste, Maurice Goursat, and Gabriel Ruget. Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications. *IEEE Transactions on Automatic Control*, 25(3):385–399, 1980.

[3] Giulio Bottegal, Aleksandr Y. Aravkin, Høakan Hjalmarsson, and Gianluigi Pillonetto. Robust em kernel-based methods for linear system identification. *Automatica*, 67:114–126, 2016.

[4] Olivier Cappé, Arnaud Doucet, Marc Lavielle, and Eric Moulines. Simulation-based methods for blind maximum-likelihood filter identification. *Signal Processing*, 73(1-2):3–25, 1999.

[5] Francesco Dinuzzo. Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, 53(5):3299–3317, 2015.

[6] Georgios B. Giannakis and Jerry M. Mendel. Identification of nonminimum phase systems using higher order statistics. *IEEE Transactions on Signal Processing*, 37(3):360–377, 1989.

[7] John J. Kormylo and Jerry M. Mendel. Identifiability of nonminimum phase linear stochastic systems. *IEEE Transactions on Automatic Control*, 28(12):1081–1090, 1983.

[8] K.S. Lii and M. Rosenblatt. Deconvolution and estimation of transfer function phase and coefficients for nongaussian linear processes. *The Annals of Statistics*, 10(4):1195–1208, 1982.

[9] Gianluigi Pillonetto and Giuseppe De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

[10] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[11] Jitendra K. Tugnait. Identification of nonminimum phase linear stochastic systems. In *Proceedings of the 23rd Conference on Decision and Control*, pages 342–347, 1984.