DENOISING THE BANANA
AN R VINAIGRETTE

ROGER KOENKER

Abstract. A simple bivariate Gaussian deconvolution problem is used to illustrate some recent progress involving the Kiefer-Wolfowitz NPMLE for mixture models.

Sometime in the early 1990's I made my last trip to Bell Labs in Murray Hill to give a talk in the Statistics Department. While there I spoke with Don X. Sun about a project that he had been working on to identify fruits and vegatables as they came down the conveyor belt in a supermarket checkout station. One way to look at this problem was to try to estimate the outer envelope of the object and this seemed to be something that quantile regression might be deployed to do. I continued to think about this problem periodically. It was tentatively called "How to recognize a banana." Whenever I saw a new idea about multivariate quantiles I would ask, "Can you estimate a banana?" This was a non-trivial test case since it involved a non-convex object, and required rather strong equivariance properties. I think that it is fair to say that none of the quantile regression approaches were a resounding success.

Then a few days ago I discovered a wonderful new paper, Soloff et al. (2021), that I should have seen much earlier since it appeared on arXiv last September. It provides a very thorough analysis of the NPMLE for multivariate heteroscedastic Gaussian mixtures. An example illustrating their approach appears early on in the paper: Gaussian observations with means on a circle of radius 2 are generated, and the bivariate NPMLE estimates the mixing distribution, G, and is used as a "plug-in" prior to denoise the observations delivering a very impressive posterior mean fit closely clustered around the original circle. More to the point the fit is nearly as good as an oracle fit that knows the empirical distribution of the true means. For amusement I thought I should try this on my ancient banana problem, so I coded a simplistic version of our usual REBayes style function to fit bivariate Gaussians and cooked up a test problem.

Figure 1 shows a somewhat artificial banana surrounded by some black dots. The latter are used as centers for 1000 observations appearing in Figure 2. Now, given the noisy red points we can estimate the mixing distribution and use it to compute posterior means for each of the observed red points. These shrunken (posterior mean) points are depicted in Figure 3 in blue.

```
> library(png)
> require(mvtnorm)
> R <- readPNG("banana.png")
> R <- as.raster(R)
> plot(1:10, 1:10, col = 0, xlab = "",  ylab = "")
> rasterImage(R, 1, 1, 10, 10)
```

January 8, 2022. A genre manifesto for R Vinaigrettes is available at `http://davoidofmeaning.blogspot.com/2016/12/r-vinaigrettes.html`.

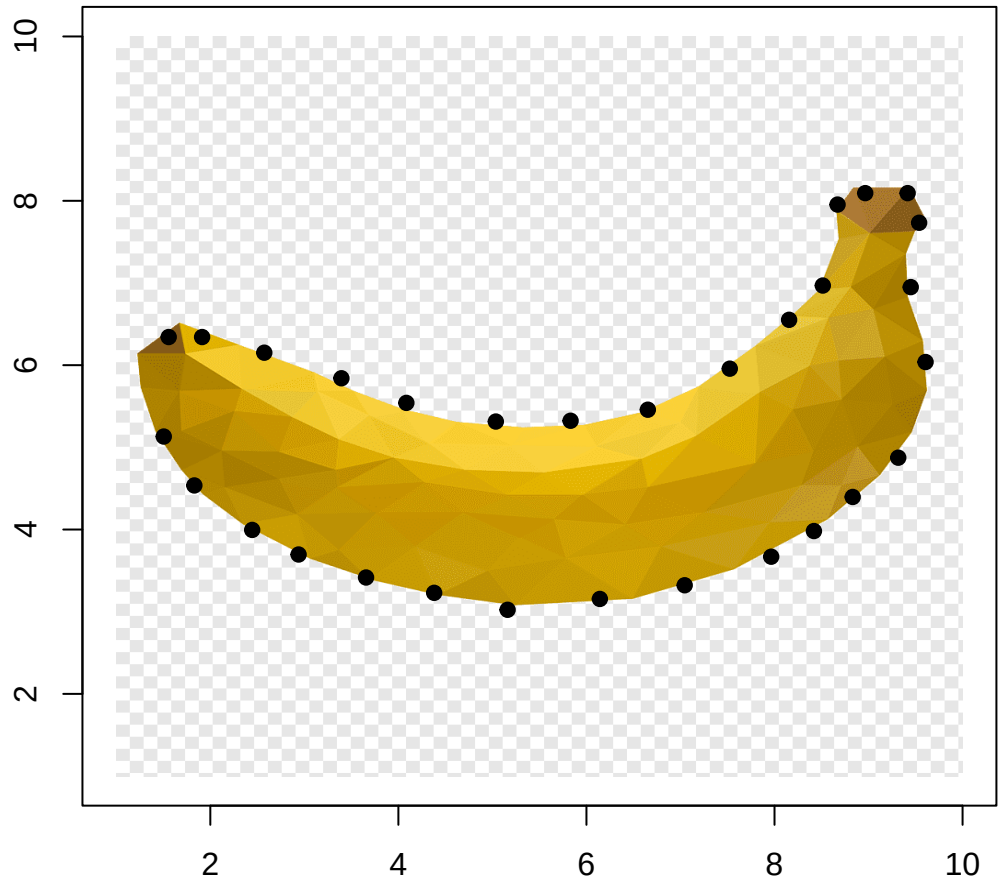

FIGURE 1. This is not a banana.

```
> load("banana.Rda") # centers via locator(30)
> points(banana, pch = 19)
```

```
> rbanana <- function(n, b = banana, sd = 0.25){
+     k <- sample(1:length(b$x), n, replace = TRUE)
+     t <- matrix(rnorm(2 * n, c(b$x[k],b$y[k]), sd = sd), n, 2)
+     dimnames(t)[[2]] <- c("x", "y")
+     class(t) <- "banana"
+     t
+ }
> B <- rbanana(1000)
```

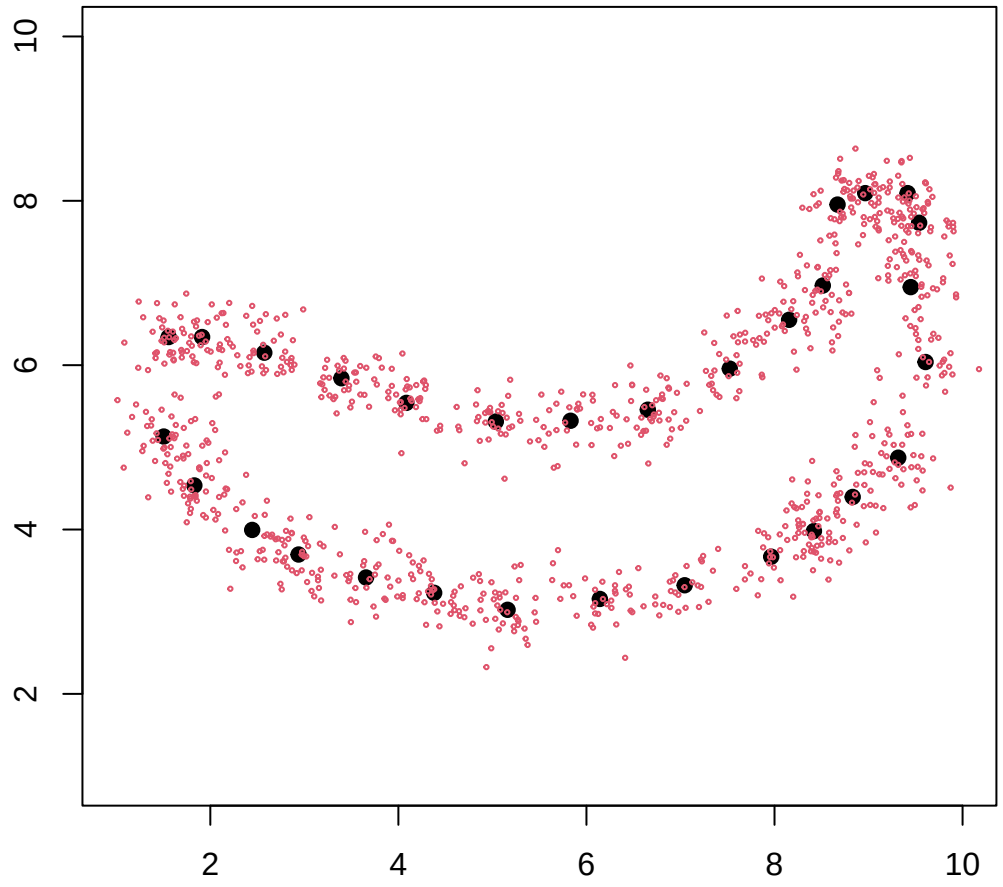

FIGURE 2. Noisy banana with centers

```
> plot(1:10, 1:10, col = 0, xlab = "", ylab = "")
> points(banana, pch = 19)
> points(B, cex = .3, col = 2)
```

```
> plot(1:10, 1:10, col = 0, xlab = "", ylab = "")
> points(banana, pch = 19)
> points(B, cex = .3, col = 2)
> points(z$du, z$dv, cex = 0.3, col = 4)
```

This is obviously just a simplistic imitation of the toy example in Soloff et al. (2021). As usual with bivariate gridding it is rather awkward to code and at least in the present

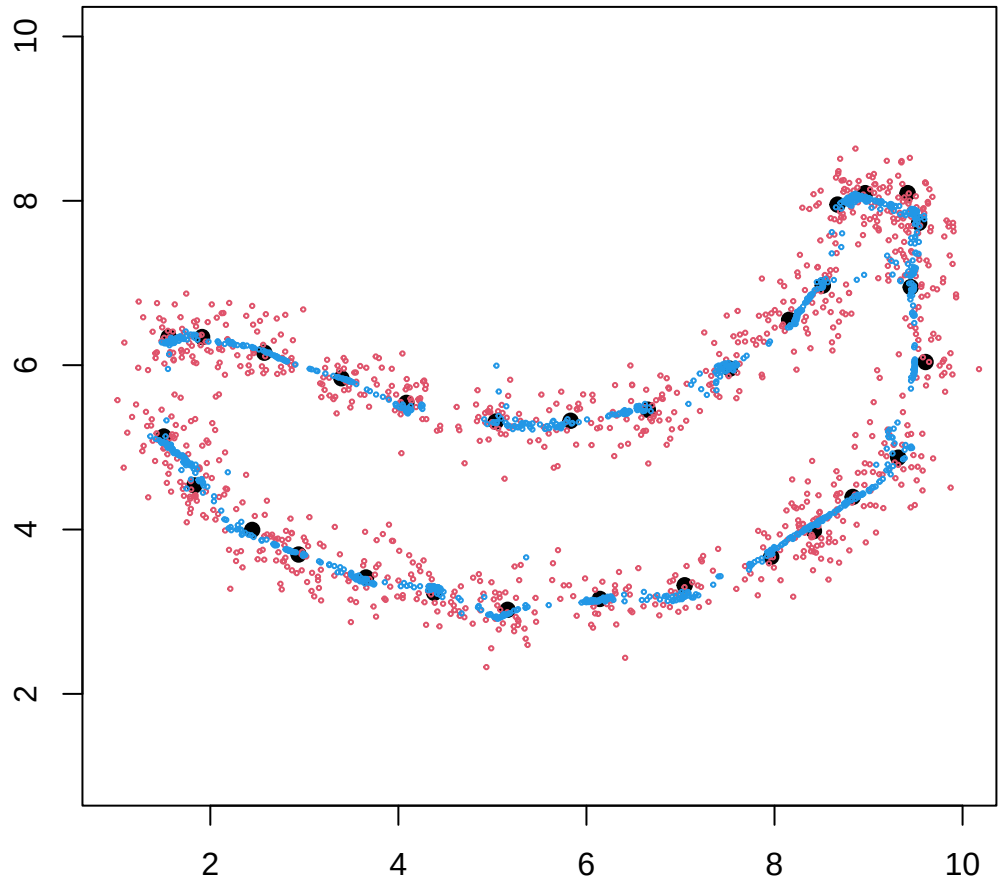

FIGURE 3. Posterior mean banana

implementation in R, quite slow. This is primarily due to the apply step to evaluate the bivariate Gaussian density, the optimization is actually very quick. The code should also be generalized to accommodate heteroscedastic noise, which has not yet been implemented in the R version. Finally, this is still not addressing the original objective of estimating the outer envelope, it only targets the conditional mean of the banana. La lutte continué.

APPENDIX A. R CODE FOR BIVARIATE GAUSSIAN MIXTURES

```
> GL2mix
function(X, S, u = 30, v = 30, eps = 1e-04, ...){
    # Should be extended to heteroscedastic S
```

```
    n <-  nrow(X)
    w <- rep(1/n,n)
    if (length(u) == 1)
        u <- seq(min(X[,1]) - eps, max(X[,1]) + eps, length = u)
    if (length(v) == 1)
        v <- seq(min(X[,2]) - eps, max(X[,2]) + eps, length = v)
    pu <- length(u)
    du <- rep(1, pu)
    pv <- length(v)
    dv <- rep(1, pv)
    UV <- expand.grid(u,v)
    A <- array(0,c(n,p,2))
    A[,,1] <- outer(X[,1], UV[,1], "-")
    A[,,2] <- outer(X[,2], UV[,2], "-")
    A <- apply(A, 1:2, function(x) dmvnorm(x, sigma = diag(2)))
    duv <- as.vector(kronecker(du, dv))
    f <- KWDual(A,duv, w, ...)
    fuv <- f$f
    uv <- expand.grid(alpha = u, theta = v)
    g <- as.vector(A %*% (duv * fuv))
    du <- A %*% (uv[, 1] * duv * fuv)/g
    dv <- A %*% (uv[, 2] * duv * fuv)/g
    z <- list(u = u, v = v, fuv = fuv, logLik = logLik, du = du,
        dv = dv, A = A, status = f$status)
    class(z) <- "GLVmix"
    z
}
```

References

Soloff, J. A., Guntuboyina, A. and Sen, B. (2021), 'Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood'. Available from `https://arxiv.org/abs/2109.03466`.