Contents

Chapter 1

Introduction and Asymptotic Analysis for Tukey Depth in the Bivariate Exponential Distribution

1.1 Introduction

In this chapter, we study the Tukey (half-space) depth for a bivariate distribution with independent exponential marginals. Let

$$X \sim \text{Exp}(\lambda_1) \quad \text{and} \quad Y \sim \text{Exp}(\lambda_2),$$

with joint density

$$f_{X,Y}(x,y) = \lambda_1 \lambda_2 \, e^{-\lambda_1 x - \lambda_2 y}, \quad x, y \geq 0.$$

The Tukey depth of a point is traditionally defined as the minimum probability mass of any closed half-space that contains the point. In our analysis, rather than simply comparing the probability mass P on one side of a fixed line with $1 - P$, we consider a family of lines (or equivalently, directions) passing through the point (a, b) and study the probability mass on one side of each line. Although one may be tempted to express the depth as

$$D(a,b) = \min\{P, 1-P\},$$

this formulation is not a complete definition of Tukey depth. First, it applies only for a given direction; the full depth is obtained by taking the infimum over all directions. Second, while for points with large coordinates (i.e. $x \gg 0$ and $y \gg 0$) one side of the half-space tends to dominate, for points near the origin the "inner" probability (the mass on the side closer to the high-density region) can actually be lower than the "outer" probability. Hence, one cannot simply compute $1-P$ and then take the minimum; the correct notion of depth requires a careful examination of the probability mass as a function of the line parameters and then taking the infimum over all admissible directions.

1.2 Definition of the Probability Mass $I(k)$

Consider a line through (a, b) given by

$$y - b = k\,(x - a), \tag{1.1}$$

with $k = \tan\theta$. In the first quadrant ($x \geq 0$, $y \geq 0$), the half-space defined by this line generally leads to different integration regions depending on where the line crosses the axes. Figure 1.1 illustrates three cases:

Case I (Triangle): Both the x- and y-intercepts are positive. In this case, the line crosses the x-axis at

$$x_0 = a - \frac{b}{k} > 0,$$

and the y-axis at

$$y_0 = b - ka > 0.$$

Hence, for $x \in [0, x_0]$ the integration in y runs from 0 up to

$$y_{\max}(x) = k(x - a) + b,$$

and the region is the simple triangle formed by the line and the axes.

Case II (Negative x-intercept, Positive y-intercept): If the x-intercept is negative, i.e.,

$$x_0 = a - \frac{b}{k} < 0,$$

while the y-intercept remains positive, then the region in the first quadrant is defined by $x \geq 0$ and $y \in \big[0,\ k(x - a) + b\big]$.

Case III (Positive x-intercept, Negative y-intercept): If the line crosses the y-axis in the negative region,

$$y_0 = b - ka < 0,$$

while $x_0 = a - \frac{b}{k} > 0$, then the boundary relevant in the first quadrant starts at $x = x_0$ (since for $x < x_0$ the line lies below $y = 0$) and extends for $x \geq x_0$ with y between 0 and $k(x - a) + b$.

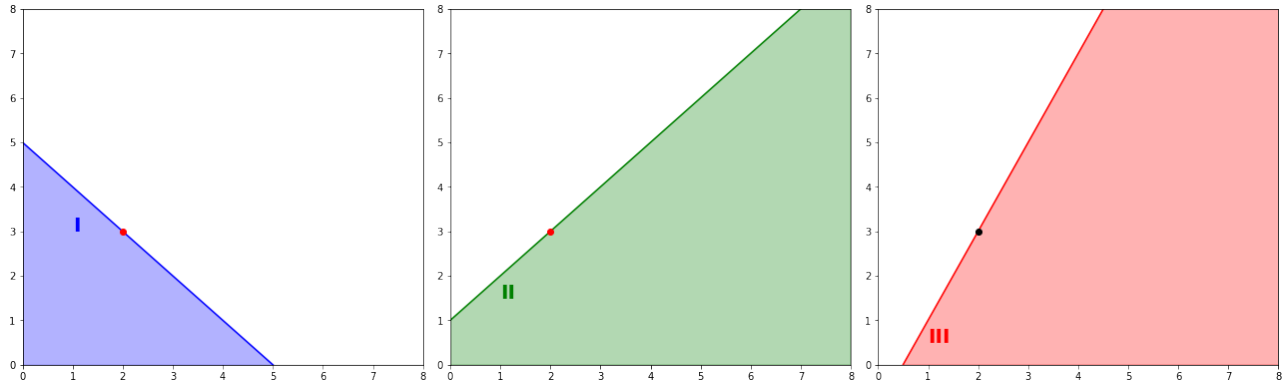

Figure 1.1: Visualization of Regions Defined by Lines Through (a, b): Case I shows the triangular region (both intercepts positive), Case II the region when the x-intercept is negative, and Case III when the y-intercept is negative.

In fact, Cases II and III can be grouped together. In these cases, the integration region in the first quadrant is given by

$$x \geq \max\{0, x_0\} \quad \text{and} \quad 0 \leq y \leq k(x - a) + b.$$

Thus, a unified expression for the probability mass below the line is

$$I(k) = \int_{x=\max\{0,x_0\}}^{\infty} \int_{y=0}^{k(x-a)+b} \lambda_1 \lambda_2 \, e^{-\lambda_1 x - \lambda_2 y} \, dy \, dx, \tag{1.2}$$

with the understanding that in Case I (the triangle) the integration naturally runs over $x \in [0, x_0]$.

1.3 Analytical Computation of $I(k)$

To identify the optimal slope k^* minimizing the half-space probability mass, we need to differentiate $I(k)$ with respect to k and solve $dI/dk = 0$. As discussed, the integration domain depends on the intercepts, so we consider the two scenarios separately.

Case I (Triangle Region)

When both intercepts are positive the integration region is

$$x \in [0, x_0], \quad \text{with} \quad x_0 = a - \frac{b}{k} > 0.$$

In this case, the probability mass is given by

$$I_1(k) = \int_0^{x_0} \lambda_1 e^{-\lambda_1 x} \left[1 - e^{-\lambda_2 \left(k(x-a)+b \right)} \right] dx. \tag{1.3}$$

This integral can be computed in closed form. In fact, writing

$$I_1(k) = \underbrace{\int_0^{x_0} \lambda_1 e^{-\lambda_1 x} \, dx}_{=1-e^{-\lambda_1 x_0}} - \underbrace{\int_0^{x_0} \lambda_1 e^{-\lambda_1 x} e^{-\lambda_2 \left(k(x-a)+b \right)} dx}_{:=J(k)},$$

we note that

$$J(k) = e^{-\lambda_2(ka-b)} \int_0^{x_0} \lambda_1 e^{-(\lambda_1+\lambda_2 k)x} \, dx = e^{-\lambda_2(ka-b)} \, \frac{\lambda_1}{\lambda_1 + \lambda_2 k} \left[1 - e^{-(\lambda_1+\lambda_2 k)x_0} \right].$$

Thus, we obtain

$$I_1(k) = \left[1 - e^{-\lambda_1 \left(a - \frac{b}{k} \right)} \right] - e^{-\lambda_2(ka-b)} \, \frac{\lambda_1}{\lambda_1 + \lambda_2 k} \left[1 - e^{-(\lambda_1+\lambda_2 k)\left(a - \frac{b}{k} \right)} \right].$$

This expression, though somewhat involved, is fully analytical and can be differentiated with respect to k using standard calculus rules.

Cases II+III (Grouped Region)

When one of the intercepts is non-positive, the integration region in the first quadrant becomes

$$x \geq \max\{0, x_0\}, \quad \text{with} \quad x_0 = a - \frac{b}{k}.$$

Define

$$A = \max\Big\{0,\, a - \frac{b}{k}\Big\}.$$

Then, the probability mass is given by

$$I_2(k) = \int_A^\infty \lambda_1 e^{-\lambda_1 x}\Big[1 - e^{-\lambda_2\left(k(x-a)+b\right)}\Big]dx. \tag{1.4}$$

As in Case I, we split the integral:

$$I_2(k) = \left[\int_A^\infty \lambda_1 e^{-\lambda_1 x} dx\right] - \left[\int_A^\infty \lambda_1 e^{-\lambda_1 x} e^{-\lambda_2\left(k(x-a)+b\right)} dx\right].$$

The first term is elementary:

$$\int_A^\infty \lambda_1 e^{-\lambda_1 x} dx = e^{-\lambda_1 A}.$$

For the second term, we rewrite

$$e^{-\lambda_2\left(k(x-a)+b\right)} = e^{-\lambda_2 k(x-a)}\, e^{-\lambda_2 b} = e^{\lambda_2 ka - \lambda_2 b}\, e^{-\lambda_2 kx}.$$

Thus,

$$\int_A^\infty \lambda_1 e^{-\lambda_1 x} e^{-\lambda_2\left(k(x-a)+b\right)} dx = e^{\lambda_2 ka - \lambda_2 b} \int_A^\infty \lambda_1 e^{-(\lambda_1+\lambda_2 k)x} dx.$$

Evaluating the remaining integral, we have

$$\int_A^\infty \lambda_1 e^{-(\lambda_1+\lambda_2 k)x} dx = \frac{\lambda_1}{\lambda_1 + \lambda_2 k}\, e^{-(\lambda_1+\lambda_2 k)A}.$$

Therefore, the closed-form expression for $I_2(k)$ is

$$I_2(k) = e^{-\lambda_1 A} - \frac{\lambda_1}{\lambda_1 + \lambda_2 k}\, e^{\lambda_2 ka - \lambda_2 b - (\lambda_1+\lambda_2 k)A}.$$

This expression can be rewritten as

$$I_2(k) = e^{-\lambda_1 A}\left[1 - \frac{\lambda_1}{\lambda_1 + \lambda_2 k}\, e^{\lambda_2 k(a-A) - \lambda_2 b}\right],$$

with

$$A = \max\left\{0,\, a - \frac{b}{k}\right\}.$$

1.4 Asymptotic Behaviour: Small-Scale Regime

In the small-scale regime the closed-form expressions for the probability mass corresponding to a half-space defined by a line through (a, b) simplify considerably. Recall that we derived

$$I_1(k) = \Big[1 - e^{-\lambda_1\left(a - \frac{b}{k}\right)}\Big] - e^{-\lambda_2(ka-b)}\, \frac{\lambda_1}{\lambda_1 + \lambda_2 k}\Big[1 - e^{-(\lambda_1+\lambda_2 k)\left(a - \frac{b}{k}\right)}\Big],$$

for the triangular region (where both intercepts are positive, with integration over

$$x \in \Big[0, a - \frac{b}{k}\Big],$$

and

$$I_2(k) = e^{-\lambda_1 A}\left[1 - \frac{\lambda_1}{\lambda_1 + \lambda_2 k}\, e^{\lambda_2 k(a-A)-\lambda_2 b}\right],$$

for the grouped region (where at least one intercept is non-positive, with

$$x \in \left[\max\left\{0, a - \frac{b}{k}\right\}, \infty\right),$$

and we define

$$A = \max\left\{0,\ a - \frac{b}{k}\right\}.$$

In our small-scale analysis we assume that the argument

$$g(x) = k(x - a) + b$$

remains sufficiently small so that a first-order Taylor expansion is valid. With the standard expansion

$$e^{-z} \approx 1 - z,$$

we expand the exponentials in the above expressions. After some algebra, one finds that the optimal slope (minimizing the half-space measure) satisfies

$$k^* \approx \frac{\lambda_1}{\lambda_2} \quad \text{or equivalently} \quad \tan\theta^* \approx \frac{\lambda_1}{\lambda_2}.$$

Assuming that

$$a - \frac{b\lambda_2}{\lambda_1} > 0,$$

we then have, for the grouped region, an effective lower integration limit

$$A^* = a - \frac{b\lambda_2}{\lambda_1}.$$

Under this condition the algebra shows that the expression corresponding to the grouped region becomes asymptotically lower. In particular, substituting $k^* = \lambda_1/\lambda_2$ yields

$$I_2(k^*) \approx \frac{1}{2}\, e^{-\lambda_1\left(a - \frac{b\lambda_2}{\lambda_1}\right)}.$$

Thus, in the regime where

$$a - \frac{b\lambda_2}{\lambda_1} > 0,$$

the infimum over the half-space measures (i.e. the Tukey depth) is asymptotically approximated by the value obtained in the grouped region.

1.5 Summary

In this chapter we developed an analytical framework for approximating the infimum of the half-space measure for a bivariate exponential distribution. We derived closed-form expressions for the probability mass $I(k)$ associated with half-spaces defined by lines through (a, b) in two cases:

- **Case I (Triangular Region):** When both intercepts are positive, with integration over

$$x \in \Big[0, a - \frac{b}{k}\Big],$$

- **Cases II+III (Grouped Region):** When at least one intercept is non-positive, with integration over

$$x \in \Big[\max\Big\{0, a - \frac{b}{k}\Big\}, \infty\Big).$$

Focusing on the small-scale regime (i.e. when $g(x) = k(x - a) + b$ is small) and applying a first-order Taylor expansion, we found that the optimal slope satisfies

$$k^* \approx \frac{\lambda_1}{\lambda_2} \quad (\tan\theta^* \approx \lambda_1/\lambda_2).$$

Assuming further that

$$a - \frac{b\lambda_2}{\lambda_1} > 0,$$

the analysis shows that the half-space corresponding to the grouped region provides the lower measure, and its value is asymptotically given by

$$I_2(k^*) \approx \frac{1}{2}\, e^{-\lambda_1\left(a - \frac{b\lambda_2}{\lambda_1}\right)}.$$

This expression represents the infimum over all half-spaces (and hence the Tukey depth) in the considered regime. Future work will address the complementary case where

$$a - \frac{b\lambda_2}{\lambda_1} \leq 0,$$

and will compare these asymptotic predictions with full numerical evaluations.

Chapter 2

Projection of an Exponential Point and Hypoexponential Distribution

In this chapter we consider the same joint density

$$f_{X_1,X_2}(x_1,x_2) = \lambda_1\lambda_2\, e^{-\lambda_1 x_1 - \lambda_2 x_2}, \quad x_1, x_2 \geq 0,$$

and analyze the depth of a fixed point (a, b) in R^2. Here, the random vector is denoted by

$$\mathbf{X} = (X_1, X_2),$$

and we study the projection of $\mathbf{X}$ onto an arbitrary direction.

2.1 Projection onto a Given Direction and Half-Space Definition

Let $u = (\cos\theta, \sin\theta)$ be an arbitrary unit vector. The projection of the random vector $\mathbf{X} = (X_1, X_2)$ onto u is given by

$$Z = u^T\mathbf{X} = X_1\cos\theta + X_2\sin\theta.$$

For a fixed point (a, b), its projection is

$$u^T(a,b) = a\cos\theta + b\sin\theta.$$

Define the half-space associated with the direction u as

$$H(u) = \{x \in R^2 :\, u^T x \geq u^T(a,b)\}.$$

Then the Tukey depth of (a, b) is defined as

$$D(a,b) = \inf_{u\in S^1}\Big\{P\big(x \in R^2 :\, u^T x \geq u^T(a,b)\big)\Big\} = \inf_{u\in S^1}\Big\{1 - F\big(u^T(a,b)\big)\Big\},$$

where F is the cumulative distribution function (CDF) of the projection Z.

2.2 Projection and the Hypoexponential Distribution

Since X_1 and X_2 are independent exponential random variables, the projection

$$Z = X_1 \cos\theta + X_2 \sin\theta$$

can be viewed as a sum of two independent scaled exponentials. Define

$$\mu_1 = \frac{\lambda_1}{\cos\theta} \quad \text{and} \quad \mu_2 = \frac{\lambda_2}{\sin\theta}.$$

If $\mu_1 \neq \mu_2$, then Z follows a hypoexponential distribution with CDF

$$F(z) = 1 - \frac{\mu_2}{\mu_2 - \mu_1} e^{-\mu_1 z} + \frac{\mu_1}{\mu_2 - \mu_1} e^{-\mu_2 z}, \quad z \geq 0.$$

In the special case when $\mu_1 = \mu_2 = \mu$ (which occurs if $\lambda_1 \sin\theta = \lambda_2 \cos\theta$), the CDF becomes

$$F(z) = 1 - (1 + \mu z)e^{-\mu z}.$$

Thus, the probability that $\mathbf{X}$ falls in the half-space is

$$P\{u^T \mathbf{X} \geq u^T(a,b)\} = 1 - F\big(u^T(a,b)\big),$$

and the depth is

$$D(a,b) = \inf_{u \in S^1} \Big\{ 1 - F\big(a\cos\theta + b\sin\theta\big) \Big\}.$$

2.3 Summary and Optimization Approach

The key idea is that the projection $Z = X_1 \cos\theta + X_2 \sin\theta$ follows a hypoexponential distribution (with the special case $\mu_1 = \mu_2$ considered separately). Consequently, the probability mass in the half-space determined by u is

$$1 - F\big(a\cos\theta + b\sin\theta\big),$$

and the Tukey depth of (a,b) is obtained by taking the infimum of this expression over all directions u (or equivalently, over θ). In practice, one may evaluate

$$1 - F\big(a\cos\theta + b\sin\theta\big)$$

explicitly using the hypoexponential CDF formulas provided above and then use numerical methods (such as grid search or optimization routines) to find

$$D(a,b) = \inf_{\theta \in [0, 2\pi)} \Big\{ 1 - F\big(a\cos\theta + b\sin\theta\big) \Big\}.$$

This formulation reduces the multidimensional depth problem to a one-dimensional optimization problem involving the hypoexponential CDF.