Statistica Sinica **19** (2009), 301-314

ON THE UNIFIED METHOD WITH NUISANCE PARAMETERS

Bodhisattva Sen, Matthew Walker and Michael Woodroofe

The University of Michigan

Abstract: In this paper we consider the problem of constructing confidence intervals in the presence of nuisance parameters. We discuss a generalization of the unified method of Feldman and Cousins (1998) with nuisance parameters. We demonstrate our method with several examples that arise frequently in High Energy Physics and Astronomy. We also discuss the hybrid resampling method of Chuang and Lai (1998, 2000), and implement it in some of the problems.

Key words and phrases: Confidence intervals, EM algorithm, hybrid resampling method, mixture distribution, profile likelihood, signal and noise.

1. Introduction

Confidence regions consisting of parameter values with high relative likelihood have a long tradition in Statistics and have generated a large literature, much of which emphasizes asymptotic calculations. See Reid (2003) for a recent survey article and Reid and Fraser (2003) for a relevant application. In an influential paper, Feldman and Cousins (1998) showed how to implement construction with exact coverage probabilities in problems, with moderate sample sizes and boundary effects, like a positive normal mean or a Poisson rate that is known to exceed a background value, that are of interest in High Energy Physics. They called the construction the unified method because it makes a natural transition from a one-sided confidence bound to a two-sided confidence interval. This method has since attracted wide interest among high energy physicists, see Mandelkern (2002). Only problems without nuisance parameters were considered in Feldman and Cousins (1998). Here we retain the interest in problems with boundary effects and moderate sample sizes, but focus on problems with nuisance parameters in addition to the parameter of primary interest.

To describe the unified method and understand the issues, suppose that a data vector X has a probability density (or mass function, in the discrete case) $f_{\theta,\eta}$ where θ is the parameter of interest and η is a nuisance parameter. For example, if a mass θ is measured with normally distributed error with an unknown standard deviation, then θ is of primary interest and the standard

deviation of the measurement is a nuisance. Let L denote the likelihood function, i.e., $L(\theta, \eta|x) = f_{\theta,\eta}(x)$; further, let $\hat{\eta}_\theta = \hat{\eta}_\theta(x)$ be the value of η that maximizes $L(\theta, \eta|x)$ for a fixed θ; let $\hat{\theta} = \hat{\theta}(x)$ and $\hat{\eta} = \hat{\eta}(x)$ be the values of θ and η that maximize $L(\theta, \eta|x)$ over all allowable values; and let

$$\Lambda_\theta(x) = \frac{L(\theta, \hat{\eta}_\theta(x)|x)}{L(\hat{\theta}(x), \hat{\eta}(x)|x)}. \tag{1.1}$$

Then unified confidence intervals consist of θ for which $\Lambda_\theta(x) \geq c_\theta$, where c_θ is a value whose computation is discussed below.

For a desired level of coverage $1 - \alpha$, a literal (and correct) interpretation of "confidence" requires that $P_{\theta,\eta}[\Lambda_\theta(X) \geq c_\theta] \geq 1 - \alpha$ for all θ and η, where $P_{\theta,\eta}$ denotes probability computed under the assumption that the parameter values are θ and η. Equivalently it requires $\min_\eta P_{\theta,\eta}[\Lambda_\theta(X) \geq c_\theta] \geq 1 - \alpha$ for each θ. Thus, c_θ should be the largest value of c for which

$$\min_\eta P_{\theta,\eta}\big[\Lambda_\theta(X) \geq c\big] \geq 1 - \alpha. \tag{1.2}$$

For a fixed x, the confidence interval is then $\mathcal{C}(x) = \{\theta : \Lambda_\theta(x) \geq c_\theta\}$, and its coverage probability

$$P_{\theta,\eta}\big[\theta \in \mathcal{C}(X)\big] = P_{\theta,\eta}\big[\Lambda_\theta(X) \geq c_\theta\big] \geq 1 - \alpha, \tag{1.3}$$

by construction. Being likelihood based, unified confidence intervals are generally reliable, even optimal, in large samples, but not necessarily so in small samples, and unified confidence intervals have been criticized in that context – e.g., Roe and Woodroofe (1999, 2000).

In some simple cases, it is possible to compute c_θ analytically. This is illustrated in Sections 2 and 3. In other cases, one can in principle proceed by numerical calculation. This requires computing $P_{\theta,\eta}[\Lambda_\theta(X) \geq c]$ over a grid of (θ, η, c) values, either by Monte-Carlo or numerical integration, and then finding the c_θ by inspection, replacing the minimum in (1.2) by the minimum over the grid. This is feasible if η is known or absent, and was done by Feldman and Cousins in two important examples. But if η is present and unknown, then numerical calculations become unwieldy, especially if η is a vector.

One way to circumvent the unwieldy numerical problems when η is present, is to use the chi-squared approximation to the distribution of Λ_θ, as in Rolke, López and Conrad (2005), or a chi-squared approximation supplemented by a Bartlett correction. Another is to use the hybrid resampling method of Chuang and Lai (1998, 2000). We generate random variable X^* from $P_{\theta,\hat{\eta}_\theta}$ and let $c_\theta^+ = c_\theta^+(x)$ be the largest values of c for which $P_{\theta,\hat{\eta}_\theta}[\Lambda_\theta(X^*) \geq c] \geq 1 - \alpha$.

Then the hybrid confidence intervals consist of θ for which $\Lambda_\theta(x) \geq c_\theta^+$. This requires computation over a grid of θ values, but not over η for fixed θ. Unfortunately, (1.3) cannot be asserted for the hybrid intervals, but Chuang and Lai argue both theoretically and by example that it should be approximately true. In some cases the calculations can be done by numerical integration, but they can always be done by simulation. For a given x, generate independent $X_1^*, \ldots, X_N^*$ (pseudo) random numbers from the density $f_{\theta,\hat{\eta}_\theta}$; compute $\Lambda_\theta(X_k^*)$ from (1.1) with x replaced by X_k^*; and let c_θ^* be the largest value of c for which

$$\frac{\#\{k \leq N : \Lambda_\theta(X_k^*) \geq c\}}{N} \geq 1 - \alpha. \tag{1.4}$$

Here the left side of (1.4) provides a Monte Carlo Estimate for $P_{\theta,\hat{\eta}_\theta}[\Lambda_\theta(X^*) \geq c]$, and c_θ^* provides an estimate of c_θ^+.

The hybrid method resembles Efron's bootstrap resampling method, but differs in one important respect: computing (1.2) for fixed θ, θ and η are replaced by θ and $\hat{\eta}_\theta$, as opposed to $\hat{\theta}$ and $\hat{\eta}$. This is the origin of the term "hybrid". Evidence that the hybrid method is reliable – that is, that (1.3) is approximately true comes from two sources, asymptotic approximations and simulations. These are reported in Chuang and Lai (1998, 2000), and include some dramatic successes. Here the method is applied to three examples of interest to astronomers and physicists. The hybrid method has (independently) been suggested in the physics literature by Feldman (2000).

Section 2 describes the analytic computation of c_θ based on (1.2) for a normal model with mean $\theta \geq 0$ and unknown variance σ^2 (σ^2 is the nuisance parameter). In Section 3 we work out the details of the method when the parameter of interest is the angle between the mean vector of a bivariate normal population. This example has applications in Astronomy. The third example we look at is a version of the "signal plus noise" problem that arises often in High Energy Physics. We observe $N \sim \text{Poisson}(b + \theta)$ and independently $M \sim \text{Poisson}(\gamma b)$, where γ is a known constant; here θ is the signal rate (the parameter of interest) and b is the background rate (a nuisance parameter). The aim is to construct a $1 - \alpha$ confidence interval for θ. We are not able to analytically compute c_θ for this example. The details are provided in Section 4. An extension of this problem is treated in Section 5 with an application to Astronomy. With every "event" we also observe a random variable with distribution depending on the type of "event" (signal event or background event). We use the EM algorithm to maximize the likelihood of this mixture model. We construct a $1 - \alpha$ confidence interval for θ using the hybrid resampling method. This generalization also arises in High Energy Physics. As will become clear in Section 2, there can be a large

difference between the literal interpretation of confidence and the hybrid method approximation. It would be interesting to understand this difference in more detail.

2. The Normal Case

Suppose that $X = (Y, W)$, where Y and W are independent, Y is normally distributed with mean $\theta \geq 0$ and variance σ^2, and W/σ^2 has a chi-squared distribution with r degrees of freedom. For example, if data originally consists of a sample $Y_i = \theta + \epsilon_i,\ i = 1, \ldots, n$, where ϵ_i's are independent and identically distributed $N(0, \sigma^2)$, then one can let $Y = \bar{Y}$ and $W = (n-1)V^2/n$ where $\bar{Y}$ and V^2 denote the sample mean and variance of $Y_1, \ldots, Y_n$. The unknown parameters here are $\theta \geq 0$ and $\sigma^2 > 0$. Thus, the likelihood function is

$$L(\theta, \sigma^2|y, w) = \frac{1}{\sqrt{2^{r+1}\pi}\Gamma(r/2)} \frac{w^{\frac{1}{2}r-1}}{\sigma^{r+1}} \exp\left\{-\frac{1}{2\sigma^2}[(y-\theta)^2 + w]\right\}.$$

For a given θ, L is maximized by

$$\hat{\sigma}^2_\theta = \frac{1}{r+1}\big[w + (y-\theta)^2\big];$$

L is maximized with respect to θ and σ^2 jointly by $\hat{\theta} = \max[0, y] = y_+$, say, and

$$\hat{\sigma}^2 = \frac{1}{r+1}\big[w + (y_-)^2\big],$$

where $y_- = -\min[0, y]$. After some simple algebra,

$$\log[\Lambda_\theta] = -\frac{1}{2}(r+1)\log(\frac{\hat{\sigma}^2_\theta}{\hat{\sigma}^2}) = -\frac{1}{2}(r+1)\log\left[\frac{W + (Y-\theta)^2}{W + (Y_-)^2}\right].$$

Let

$$U = \frac{W}{\sigma^2} \qquad \text{and} \qquad Z = \frac{Y-\theta}{\sigma}.$$

Then U and Z are independent random variables for which $U \sim \chi^2_r$, $Z \sim$ Normal$(0, 1)$, and

$$\log[\Lambda_\theta] = -\frac{1}{2}(r+1)\log\left[\frac{U + Z^2}{U + [(Z + \theta/\sigma)_-]^2}\right].$$

This is an increasing function of σ for each $\theta > 0$. So, since the joint distribution of U and Z does not depend on parameters,

$$\min_{\sigma>0} P_{\theta,\sigma}\big[\Lambda_\theta \geq c\big] = \lim_{\sigma\to 0} P_{\theta,\sigma}\big[\Lambda_\theta \geq c\big] = P\left[-\frac{1}{2}(r+1)\log\big(1 + \frac{T^2}{r}\big) \geq \log(c)\right],$$

where $T = Z/\sqrt{U/r}$ has t-distribution with r degrees of freedom. Thus the desired c is

$$c = \exp\Big\{ -\frac{1}{2}(r+1)\log\big[1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}\big]\Big\},$$

where $t_{r,1-\alpha/2}$ is the $1-\alpha/2$ percentile of the latter distribution and is independent of θ. To find the confidence intervals, one must solve the inequality $\Lambda_\theta \geq c$ for θ. Letting $s^2 = W/r$, this may be written

$$\frac{1 + (y-\theta)^2/(rs^2)}{1 + y_-^2/(rs^2)} \leq 1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r},$$

or

$$[y - bs]_+ \leq \theta \leq y + bs, \tag{2.1}$$

where

$$b = \sqrt{t^2_{r,1-\frac{1}{2}\alpha} + \frac{y_-^2}{s^2}\left(1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}\right)}. \tag{2.2}$$

Thus, if $y > 0$, the unified intervals are just the usual t-intervals, truncated to non-negative values, and if $y > bs$ they are symmetric about y. This differs from the case of known σ, where the intervals are (slightly) asymmetric, even for large y. There is a more dramatic difference with the case of known σ for $y < 0$. Observe that for $y < 0$,

$$y + bs \geq s\sqrt{\frac{y^2}{s^2}\left(1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}\right)} - \frac{|y|}{s} = |y|\left\{\sqrt{1 + \frac{t^2_{r,1-\frac{1}{2}\alpha}}{r}} - 1\right\}.$$

So the upper confidence limit approaches $+\infty$ as $y \to -\infty$, unlike the case of known σ where it approaches 0. Mandelkern (2002) found the latter behavior non-intuitive. If we let $r \to \infty$ and $s^2 \to \sigma^2$, then we do *not* recover the intervals of Feldman and Cousins with known σ^2. Rather, we get the interval (2.1) with the t-percentile replaced by the corresponding normal percentile.

Observe that the confidence limits for θ may be written as $[y/s - b]_+ \leq \theta/s \leq y/s + b$. Figure 2.1 shows these upper and lower confidence limits for θ/s as a function of y/s for $r = 10$ and $\alpha = 0.10$. For a specific example, suppose that $r = 10$, $s = 1$, $y = -0.30$ and $\alpha = 0.10$. Then $b = \sqrt{(1.812)^2 + (0.3)^2\{1 + (1.812)^2/10\}} = 1.84$, and the interval is $0 \leq \theta \leq 1.54$. The hybrid method yields $0 \leq \theta \leq 1.14$ in this example. The details are omitted here, but an example using the hybrid method is included in Section 4.

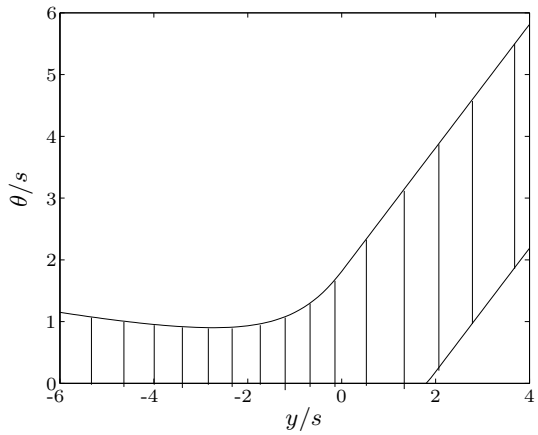

Figure 2.1. Confidence limits for θ/s as a function of y/s when $r = 10$ and $\alpha = 0.1$. Observe that the upper limit starts to increase as y decreases for $y < 0$.

3. Angles

In Astronomy, "proper motion" refers to the angular velocity of an object in the plane perpendicular to the line of sight. An object's proper motion is given by $X = (X_1, X_2)$, where X_1 and X_2 are orthogonal components and are measured independently. In certain applications astronomers are more concerned with the direction than the magnitude of the proper motion vector. An example is the motion of a satellite galaxy whose stellar orbits may be disrupted by the tidal influence exerted by a larger parent system. Due to outward streaming of its stars, a disrupting satellite will elongate spatially and exhibit a radial velocity gradient along the direction of elongation. N-body simulations indicate that the orientations of both the elongation and velocity gradient correlate with the direction of the satellite's proper motion vector (e.g., Oh, Lin and Aarseth (1995) and Piatek and Pryor (1995)). Constraining the direction of the satellite's proper motion can therefore help determine whether or not a satellite is undergoing disruption, which in turn places constraints on applicable dynamical models.

Suppose X_1 and X_2 are normally distributed random variables with unknown means μ_1 and μ_2 and known variance σ^2. Write μ_1 and μ_2 in polar coordinates as $\mu_1 = \rho\cos(\theta)$ and $\mu_2 = \rho\sin(\theta)$, where $-\pi < \theta \leq \pi$. We consider confidence intervals for θ when ρ is the nuisance parameter.

In this example, the likelihood function

$$L(\theta, \rho|x) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[(x_1 - \rho\cos(\theta))^2 + (x_2 - \rho\sin(\theta))^2\right]\right\}$$

is maximized for a fixed θ by $\hat{\rho}_\theta = \max[0, x_1\cos(\theta) + x_2\sin(\theta)]$, and unconditionally by $\hat{\rho}$ and $\hat{\theta}$ where $x_1 = \hat{\rho}\cos(\hat{\theta})$ and $x_2 = \hat{\rho}\sin(\hat{\theta})$. Then $L(\hat{\theta}, \hat{\rho}|x) = 1/(2\pi\sigma^2)$, and

$$\Lambda_\theta = \exp\big[-\frac{1}{2\sigma^2}(\hat{\rho}^2 - \hat{\rho}_\theta^2)\big].$$

Let $Z_1 = \big[\cos(\theta)X_1 + \sin(\theta)X_2 - \rho\big]/\sigma$ and $Z_2 = \big[\sin(\theta)X_1 - \cos(\theta)X_2\big]/\sigma$. Then Z_1 and Z_2 are independent normal variables each with mean 0 and unit variance, and

$$\Lambda_\theta = \exp\big\{-\frac{1}{2}[(Z_1 + \rho)_-^2 + Z_2^2]\big\},$$

where (recall) $z_- = -\min[0, z]$, after some simple algebra. Thus, Λ_θ is an increasing function of ρ for fixed Z_1, Z_2, and θ. So, since the joint distribution of Z_1 and Z_1 does not depend on parameters

$$\min_\rho P_{\theta,\rho}\big[\Lambda_\theta \geq c\big] = \lim_{\rho\to 0} P_{\theta,\rho}\big[\Lambda_\theta \geq c\big].$$

Letting $b = -2\log(c)$, this is just

$$\begin{aligned} P\big[Z_{1,-}^2 + Z_2^2 \leq b\big] &= P\big[Z_1 \leq 0, Z_1^2 + Z_2^2 \leq b\big] + P\big[Z_1 > 0, Z_2^2 \leq b\big] \\ &= \frac{1}{2}P[\chi_1^2 \leq b] + \frac{1}{2}P[\chi_2^2 \leq b]. \end{aligned}$$

So $c = e^{-b/2}$, where b solves $P[\chi_1^2 \leq b]/2 + P[\chi_2^2 \leq b]/2 = 1 - \alpha$. For example, when $\alpha = 0.90,\ b = 3.808$.

Unified confidence intervals for θ then consist of those θ for which $\hat{\rho}^2 - \hat{\rho}_\theta^2 \leq b\sigma^2$, or equivalently $\hat{\rho}_\theta^2 \geq \hat{\rho}^2 - b\sigma^2$. Thus, if $\hat{\rho}^2 \leq b\sigma^2$, the interval consists of all values $-\pi < \theta \leq \pi$. On one hand, this simply reflects the (obvious) fact that if $\hat{\rho}$ is small there is no reliable information for estimating θ, but it also admits the following amusing paraphrase: one is $100(1-\alpha)\%$ confident of something that is certain. If $\hat{\rho}^2 > b\sigma^2$, the intervals consist of those θ for which $\hat{\rho}\cos(\theta - \hat{\theta}) \geq \sqrt{\hat{\rho}^2 - b\sigma^2}$; that is,

$$\hat{\theta} - \arccos\left(\sqrt{1 - \frac{b\sigma^2}{\hat{\rho}^2}}\right) \leq \theta \leq \hat{\theta} + \arccos\left(\sqrt{1 - \frac{b\sigma^2}{\hat{\rho}^2}}\right),$$

where $\arccos(y)$ is the unique ω for which $0 \leq \omega \leq \pi$ and $\cos(\omega) = y$, and addition is understood modulo π. Thus, there is a discontinuity in the length of the intervals as $\hat{\rho}$ passes through $b\sigma^2$: it decreases from 2π to something less than π.

Piatek et al. (2002) measured the Galactic rest-frame proper motion of the Fornax galaxy to be $(X_1, X_2) = (32, 33)$ with $\sigma = 13$ (units are in milli-arcseconds per century). Dinescu, Keeney, Majewski and Girard (2004) made a similar

measurement but observed $(X_1, X_2) = (-13, 34)$ with $\sigma = 16$. We use our method to construct a 90% confidence interval for the direction θ in the two cases. The intervals obtained are $(0.2219, 1.4119)$ for the Piatek et al. angle and $(0.9051, 2.9669)$ for the Dinescu et al. angle (where θ is measured in radians). Note that the Piatek et al. measurement places a tighter constraint on the proper motion direction, and that there is some overlap with the Dinescu et al. result.

4. Counts with Background

Suppose that $X = (N, M)$ where N and M are independent, M has the Poisson distribution with mean γb, and N has the Poisson distribution with mean $b + \theta$. It is useful to write $N = B + S$ where B and S are independent Poisson random variables with means b and θ, representing the number of background and signal events. Here b and θ are unknown; γ is assumed known, and large values of γ are of interest. In this case, the likelihood function and score functions are

$$\begin{aligned} L(\theta, b|n, m) = f_{\theta,b}(n, m) &= \frac{(\gamma b)^m}{m!} e^{-\gamma b} \times \frac{(\theta + b)^n}{n!} e^{-(\theta+b)}, \\ \frac{\partial \log(L)}{\partial \theta} &= \frac{n}{b + \theta} - 1, \\ \frac{\partial \log(L)}{\partial b} &= \frac{m}{b} + \frac{n}{\theta + b} - (\gamma + 1). \end{aligned}$$

Consider $\hat{b}_\theta$ for a fixed θ. If $m = 0$, then L is maximized when $b = [n/(\gamma + 1) - \theta]_+$; if $m > 0$ it is maximized at the (positive) solution to $\partial \log(L)/\partial b = 0$, i.e.,

$$\hat{b}_\theta = \frac{[(m + n) - (\gamma + 1)\theta] + \sqrt{[(\gamma + 1)\theta - (m + n)]^2 + 4(\gamma + 1)m\theta}}{2(\gamma + 1)}; \qquad (4.1)$$

fortuitously, (4.1) also gives the correct answer when $m = 0$. The unconstrained maximum likelihood estimators are then $\hat{\theta}$ and $\hat{b} = \hat{b}_{\hat{\theta}}$, where $\hat{\theta}$ maximizes the profile likelihood function $L(\theta, \hat{b}_\theta|n, m)$. Considering the cases $n \le m/\gamma$ and $n > m/\gamma$ separately, shows that $\hat{\theta} = \big(n - m/\gamma\big)_+$, $\hat{b} = (m + n - \hat{\theta})/(\gamma + 1)$, and

$$\Lambda_\theta(n, m) = (\frac{\hat{b}_\theta}{\hat{b}})^m (\frac{\theta + \hat{b}_\theta}{\hat{\theta} + \hat{b}})^n \exp\big[(n + m) - (\gamma + 1)\hat{b}_\theta - \theta\big],$$

after some simple algebra.

We have been unable to find the minimizing value in (1.2) and so will use the Hybrid Resampling Method. This is best illustrated by an example. Figure 4.2

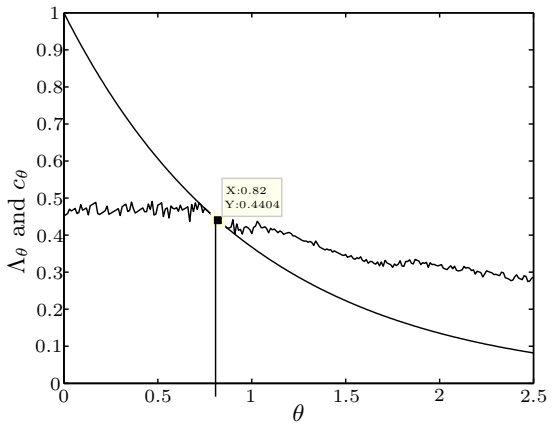

Figure 4.2. Plot of Λ_θ (smooth line) and c_θ (jagged line) against θ when $\gamma = 6, m = 23$, $n = 0$ and $\alpha = 0.10$.

below shows Λ_θ and c_θ when $\gamma = 6, m = 23$, $n = 0$, and $\alpha = 0.10$. This is patterned after the original KARMEN report Eitel and Zeitnitz (1998), but with a larger value of $\hat{b}$ and more variability in $\hat{b}$. The c^*_θ was computed by Monte Carlo on the grid $\theta = 0, 0.01, 0.02, \ldots, 2.50$ using $N = 10,000$ in (1.4). The right end-point of the interval is 0.82.

By construction, the hybrid-unified method always delivers a non-degenerate subinterval of $[0, \infty)$, even when $n = 0$, and thus it avoids the types of problems reported in Rolke, López and Conrad (2005). It does not avoid the problems inherent in the use of the unified method without nuisance parameters, however – for example, dependence of the interval on $\hat{b}$ when $n = 0$. We believe that the interval $[0, 2.31]$ is a more reasonable statement of the uncertainty in this example. Briefly, $[0, 2.31]$ would be the uniformly most accurate 90% confidence interval if $S = 0$ were observed, if $N = 0$, then $B = S = 0$.

5. The Star Contamination Problem

In studying external (to the Milky Way) galaxies, one can measure only two of the three (those orthogonal to the line of sight) components of stellar position, and one (along the line of sight, from red shift of spectral features) of the three components of stellar velocity. Because the line of sight necessarily originates within the Milky Way, velocity samples for distant galaxies frequently suffer from contamination by foreground Milky Way stars. It is important to accurately identify and remove sample contamination. The most common procedure for membership determination involves fitting a normal distribution to the marginal velocity distribution of all observed stars, then iteratively rejecting

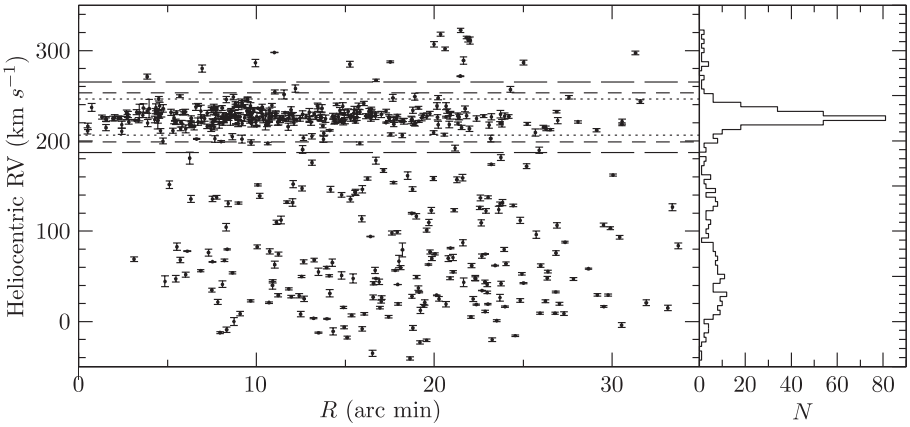

Figure 5.3. **Left:** Heliocentric radial velocities (RV) vs. angular distance from the Sextans center. Dotted, short-dashed, and long-dashed lines mark boundaries of 276, 294 and 303 member samples, respectively. **Right:** Histogram of the radial velocity of the stars.

outliers beyond a specified ($\sim 3\sigma$) threshold. However, this is of limited utility when the velocity distributions of target galaxy and contaminant stars overlap. Also, the trimming of outliers from an imposed distribution introduces a degree of circularity to the analysis, as it is the target galaxy's velocity distribution that is under investigation. We consider results from a velocity survey of the Sextans dwarf spheroidal galaxy (see Walker et al. (2006)). The unfiltered marginal velocity distribution of the 528 observed stars displays evidence of significant contamination by Milky Way foreground stars (see Figure 5.3). For the i'th star we consider the measurements $(X_{1i}, X_{2i}, U_{3i}, \sigma_i)$, where (X_{1i}, X_{2i}) is the projected position of the star, U_{3i} is the observed line-of-sight velocity, and σ_i is the error associated with the measurement of U_{3i}. In this section we develop a method of addressing sample contamination that incorporates a model of the contaminant distribution. We would like to estimate the number of "signal" (Sextans) stars and construct a $1 - \alpha$ confidence interval. Our algorithm also outputs, for each observed star, an estimate of the probability that the star belongs to the contaminant population. These probability estimates can be used as weights in subsequent analyses.

5.1. The statistical model

We assign parametric distributions to the positions and velocities of the stars; the parametric models are derived from the underlying physics in most cases. The EM algorithm is then employed to find MLE's estimates of the unknown parameters. The method is described in the context of available data, but can

be generalized to incorporate membership constraints provided by additional data (such as multi-color photometry data).

Suppose $N \sim \text{Poi}(b+\theta)$ is the number of stars observed using the telescope in a given amount of time. In our case we have $N = 528$. Here θ denotes the rate for observing a signal star, i.e., a Sextans star. We assume that the foreground rate is b. We are interested in constructing a $1-\alpha$ CI for θ. The actual line of sight velocity for the i'th star will be denoted by V_{3i}. We assume that $U_{3i} = V_{3i} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_i^2)$ and the ϵ_i's are assumed independent. Let Y_i be the indicator of a foreground star, i.e., $Y_i = 1$ if the i'th star is a foreground star, and $Y_i = 0$ otherwise. Of course, we do not observe Y_i. We need to make assumptions on the form of the joint density of $W_i = (X_{1i}, X_{2i}, U_{3i})$.

For the foreground stars (i.e., $Y_i = 1$) it might be reasonable to assume that the position (X_{1i}, X_{2i}) and velocity U_{3i} are independent. Then the joint density of W_i simplifies to $h_b(w) = f^{(b)}(x_1, x_2)g^{(b)}(u_3)$, where we take the position of the star as uniformly distributed in the field of view, i.e., $f^{(b)}(x_1, x_2) = 1/(\pi M^2)$, where M is the radius of field of view (in our data set it is 35 arc min). Note that $U_{3i} \sim g^{(b)}(\cdot)$, where $g^{(b)}$ is a completely known density obtained from the Besancón Milky Way model (Robin et al. (2003)), that specifies spatial and velocity distributions of Milky Way stars along a given line of sight. The density estimate $g^{(b)}(\cdot)$ was constructed using kernel density estimation techniques.

For the Sextans stars, there is a well-known model in Astronomy for the distribution of the projected position of stars. The model assumes $f^{(s)}(x_1, x_2) = K(h)e^{-s/h}, 0 \le s^2 = x_1^2 + x_2^2 \le M^2$, where $K(h)^{-1} = 2\pi h^2\{1 - (M/h)e^{-M/h} - e^{-M/h}\}$ is the normalizing constant (M is the radius of field of view). The distribution of U_{3i} given the position is assumed to be normal with mean μ and variance $\sigma^2 + \sigma_i^2$, and its density is denoted by $g^{(s)}(\cdot)$. Thus, the joint density of W_i given that it is a signal star is $h_{s,i}(w) = f^{(s)}(x_1, x_2)g^{(s)}(u_3)$.

5.2. CI for θ: the number of "signal" stars

The likelihood for the observed data is

$$L(\theta, \eta) = e^{-(b+\theta)} \frac{(b+\theta)^N}{N!} \prod_{i=1}^{N} \left(\frac{b h_b(W_i) + \theta h_{s,i}(W_i)}{b+\theta} \right), \tag{5.1}$$

which is a essentially a mixture density problem. A simple application of the EM algorithm (details are provided in the appendix) yields the MLE's in this scenario. The hybrid resampling method can be used to construct a confidence region for θ.

The likelihood ratio statistic is defined as in (1.1), and can be computed for each θ. The hybrid resampling method was employed to find the c_θ^+ as

described in the Introduction. Varying θ, we get a confidence interval for θ. In our example, the 90% confidence interval turns out to be $(260.3, 318.4)$. Note that if b was known, and with $\hat{\theta} \approx 290$ (the maximum likelihood estimate of θ), a 90% CI using frequentist method (obtained by intersecting uniformly most accurate 95% confidence lower and upper bounds) would be (261.7, 318.4). This shows that the hybrid method works almost as well as the optimal frequentist confidence region, even when b is unknown.

A. Appendix

We outline the implementation of the EM-algorithm, described in the last section, to find the the unconstrained maximum of the observed (incomplete) data. The constrained maximization is very similar (in fact, a bit simpler). Recall that Y_i is the indicator of a foreground star, $i = 1, \ldots, N$. Note that the Y_i's are i.i.d. Bernoulli $b/(b+\theta)$. Let $\mathbf{Z} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{U}, \mathbf{Y}, N)$ be the complete data matrix. The likelihood for the complete data can be written as

$$\tilde{L}(\theta, \eta|\mathbf{Z}) = e^{-(b+\theta)} \frac{(b+\theta)^N}{N!} \left\{ \prod_{i=1}^{N} (\frac{b}{b+\theta})^{Y_i} (\frac{\theta}{b+\theta})^{1-Y_i} h_b(W_i)^{Y_i} h_{s,i}(W_i)^{1-Y_i} \right\}.$$

The log-likelihood (up to a constant term) can be written as

$$\tilde{l}(\theta, \eta|\mathbf{Z}) = -(b+\theta) + \sum_{i=1}^{N} \{Y_i \log(bh_b(W_i)) + (1-Y_i) \log(\theta h_{s,i}(W_i))\}.$$

Letting θ_n and η_n denote the parameter values obtained in the n'th step of the iteration, the E-step in the unconstrained maximization process evaluates $E_{\hat{\theta}_n, \hat{\eta}_n} \left(\tilde{l}(\eta|\mathbf{Z})|\mathbf{W} \right)$ as

$$\begin{aligned} &\sum_{i=1}^{N} P_{\hat{\theta}_n, \hat{\eta}_n}(Y_i = 1|\mathbf{W}) \log[bh_b(W_i)] \\ &\qquad + \sum_{i=1}^{N} P_{\hat{\theta}_n, \hat{\eta}_n}(Y_i = 0|\mathbf{W}) \log[\theta h_{s,i}(W_i)] - (b+\theta), \end{aligned} \tag{A.1}$$

where $P_{\hat{\theta}_n, \hat{\eta}_n}(Y_i = 1|\mathbf{W}) = [\hat{b}_n h_b(W_i)]/[\hat{b}_n h_b(W_i) + \hat{\theta}_n h_{s,i}(W_i)]$ is the probability of a foreground star given the data under the current estimates of θ and η, i.e., θ_n and η_n. The M-step maximizes (A.1), which leads to the following estimating equations:

$$\frac{1}{b} \sum_{i=1}^{N} P_{\hat{\theta}_n, \hat{\eta}_n}(Y_i = 1|\mathbf{W}) - 1 = 0,$$

$$\frac{1}{\theta}\sum_{i=1}^{N} P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i = 0|\mathbf{W}) - 1 = 0,$$
$$\sum_{i=1}^{N} P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i = 0|\mathbf{W}) \left\{ \frac{1}{\sigma^2 + \sigma_i^2}(U_{3i} - \mu) \right\} = 0,$$
$$\sum_{i=1}^{N} P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i = 0|\mathbf{W}) \left\{ \frac{(U_{3i} - \mu)^2}{2(\sigma^2 + \sigma_i^2)^2} - \frac{1}{2(\sigma^2 + \sigma_i^2)} \right\} = 0.$$

The first two equations can be solved easily to give $\hat{b}_{n+1} = \sum_{i=1}^{N} P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i = 1|\mathbf{W})$ and $\hat{\theta}_{n+1} = \sum_{i=1}^{N} P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i = 0|\mathbf{W})$. The last two equations can be slightly modified to give the following (closed form) estimates of μ and σ^2:

$$\hat{\mu}_{n+1} = \frac{\sum_{i=1}^{N} \frac{P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i=0|\mathbf{W})}{1+\sigma_i^2/\hat{\sigma}^2_{(n)}} U_{3i}}{\sum_{i=1}^{N} \frac{P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i=0|\mathbf{W})}{1+\sigma_i^2/\hat{\sigma}^2_{(n)}}} \text{ and } \hat{\sigma}^2_{(n+1)} = \frac{\sum_{i=1}^{N} \frac{P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i=0|\mathbf{W})}{(1+\sigma_i^2/\hat{\sigma}^2_{(n)})^2} (U_{3i} - \hat{\mu}_{n+1})^2}{\sum_{i=1}^{N} \frac{P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i=0|\mathbf{W})}{1+\sigma_i^2/\hat{\sigma}^2_{(n)}}},$$

where $\hat{\sigma}^2_{(n)}$ is the n'th step estimate of σ^2. These estimates $(\hat{\eta}_n)$ stabilize after a few iterations yielding the MLE's of η with the incomplete data. An interesting feature of this solution is that at the end of the algorithm we get estimated probabilities that the i'th star is a signal star, namely, $P_{\hat{\theta}_n,\hat{\eta}_n}(Y_i = 1|\mathbf{W})$.

References

Chuang, C. and Lai, T. L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika* **85**, 317-332.

Chuang, C. and Lai, T. L. (2000). Hybrid resampling methods for confidence intervals. *Statist. Sinica* **10**, 1-50.

Dinescu, D. I., Keeney, B. A., Majewski, S. R. and Girard, T. M. (2004). *Astronomical J.* **128**, 687-699.

Eitel, K., and Zeitnitz, B. (1998). The search for neutrino oscillations $\bar{\nu}_\mu \to \bar{\nu}_e$ with KARMEN. Available at arXiv/hepex/9809007.

Feldman, G. (2000). Multiple Measurements and Parameters in the Unified Approach. Talk at the FermiLab Workshop on Confidence Limits.

Feldman, G. and Cousins, R. (1998). Unified approach to the classical statistical analysis of small signal. *Phys. Rev. D* **57**, 3873-889.

Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statist. Sci.* **17**, 149-172.

Oh, K. S., Lin, D. N. C. and Aarseth, S. J. (1995). *Astrophysical J.* **442**, 142-158.

Piatek, S. and Pryor, C. (1995). *Astronomical J.* **109**, 1071-1085.

Piatek, S., Pryor, C., Olszewski, E. W., Harris, H. C., Mateo, M., Minniti, D., Monet, D. G., Morrison, H. and Tinney, C. G. (2002). *Astronomical Journal* **124**, 3198-3221.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statis.* **31**, 1695-2095.

Reid, N. and Fraser, D. A. S. (2003). Likelihood Inference in the Presence of Nuisance Parameters. *PHYSTAT 2003, SLAC*, 265-271 (arXiv: physics/0312079).

Robin, A.C., Reylé, C., Derrière, S. and Picaud, S. (2003). *Astronomy and Astrophysics* **409**, 523-540.

Roe, B. and Woodroofe, M. (1999). Improved probability method for estimation the signal in the presence of background. *Phys. Rev. D* **60**, 053009.

Roe, B. and Woodroofe, M. (2000). Setting confidence belts. *Phys. Rev. D* **63**, 013009.

Rolke, W., López, A. and Conrad, J. (2005). Limits and Confidence Intervals in the Presence of Nuisance Parameters. *Nucl. Instrum. Meth.* A551, 493-503 (arXiv:physics/0403059 v4 7 July 2005).

Walker, M., Mateo, M., Olszewski, E., Pal, J., Sen, B. and Woodroofe, M. (2006). On Kinematic Substructure in the Sextans Dwarf Galaxy. *Astrophysical J.* **642**, L41-L44.

Department of Statistics, University of Michigan, 439 West Hall, 1085 South University, Ann Arbor, MI 48109-1092, U.S.A.

E-mail: bodhi@umich.edu

Department of Astronomy, University of Michigan, 500 Church St., 830 Dennison Ann Arbor, MI 48109-104, U.S.A.

E-mail: mgwalker@umich.edu

Department of Statistics, University of Michigan, 439 West Hall, 1085 South University, Ann Arbor, MI 48109-1092, U.S.A.

E-mail: michaelw@umich.edu

(Received October 2006; accepted July 2007)