Generalized Fiducial Inference: A Review

Jan Hannig* Hari Iyer† Randy C. S. Lai‡ Thomas C. M. Lee§

September 16, 2015

Abstract

R. A. Fisher, the father of modern statistics, proposed the idea of fiducial inference during the first half of the 20th century. While his proposal led to interesting methods for quantifying uncertainty, other prominent statisticians of the time did not accept Fisher's approach as it became apparent that some of Fisher's bold claims about the properties of fiducial distribution did not hold up for multi-parameter problems. Beginning around the year 2000, the authors and collaborators started to re-investigate the idea of fiducial inference and discovered that Fisher's approach, when properly generalized, would open doors to solve many important and difficult inference problems. They termed their generalization of Fisher's idea as generalized fiducial inference (GFI). The main idea of GFI is to carefully transfer randomness from the data to the parameter space using an inverse of a data generating equation without the use of Bayes theorem. The resulting generalized fiducial distribution (GFD) can then be used for inference.

After more than a decade of investigations, the authors and collaborators have developed a unifying theory for GFI, and provided GFI solutions to many challenging practical problems in different fields of science and industry. Overall, they have demonstrated that GFI is a valid, useful, and promising approach for conducting statistical inference. The

*Corresponding author. Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260, USA. Email: jan.hannig@unc.edu

†Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. Email: hari@nist.gov

‡Department of Mathematics and Statistics, University of Maine, 5752 Neville Hall, Room 333, Orono, ME 04469, USA. Email: chushing.lai@maine.edu

§Department of Statistics, University of California at Davis, 4118 Mathematical Sciences Building, One Shields Avenue, Davis, CA 95616, USA. Email: tcmlee@ucdavis.edu

goal of this paper is to deliver a timely and concise introduction to GFI, to present some of the latest results, as well as to list some related open research problems. It is the authors' hope that their contributions to GFI will stimulate the growth and usage of this exciting approach for statistical inference.
keywords: fiducial inference, approximate Bayesian computations, data generating equation, Jacobian calculation, model selection, uncertainty quantification

1 Introduction

The origin of fiducial inference can be traced back to R. A. Fisher (1922, 1925, 1930, 1933, 1935) who introduced the concept of a fiducial distribution for a parameter, and proposed the use of this fiducial distribution, in place of the Bayesian posterior distribution, for interval estimation of the parameter. In simple situations, especially in one parameter families of distributions, Fisher's fiducial intervals turned out to coincide with classical confidence intervals. For multi-parameter families of distributions, the fiducial approach led to confidence sets whose frequentist coverage probabilities were close to the claimed confidence levels but they were not exact in the repeated sampling frequentist sense. Fisher's proposal led to major discussions among the prominent statisticians of the mid 20th century (e.g., Dempster, 1966, 1968; Fraser, 1961a,b, 1966, 1968; Jeffreys, 1940; Lindley, 1958; Stevens, 1950; Tukey, 1957). Many of these discussions focused on the non-exactness of the confidence sets and also non-uniqueness of fiducial distributions. The latter part of the 20th century has seen only a handful of publications (Barnard, 1995; Dawid and Stone, 1982; Dawid *et al.*, 1973; Salome, 1998; Wilkinson, 1977) as the fiducial approach fell into disfavor and became a topic of historical interest only.

Since the mid 2000s, there has been a revival of interest in modern modifications of fiducial inference. This increase of interest demonstrated itself in both the number of different approaches to the problem and the number of researchers working on these problems, and is leading to an increasing number of publications in premier journals. The common thread

for these approaches is a definition of inferentially meaningful probability statements about subsets of the parameter space without the need for subjective prior information.

These modern approaches include Dempster-Shafer theory (Dempster, 2008; Edlefsen *et al.*, 2009) and recent (since 2010) related approach called *inferential models* (Martin and Liu, 2013, 2015a,b; Martin *et al.*, 2010; Zhang and Liu, 2011) that aims at provably conservative inference. A somewhat different approach termed *confidence distributions* looks at the problem of obtaining an inferentially meaningful distribution on the parameter space from a purely frequentist point of view (Xie and Singh, 2013). One of the main contributions of this approach is *fusion learning*: its ability to combine information from disparate sources with deep implications for meta analysis (Hannig and Xie, 2012; Schweder and Hjort, 2002; Singh *et al.*, 2005; Xie *et al.*, 2011, 2013). Another related approach is based on higher order likelihood expansions and implied data dependent priors (Fraser *et al.*, 2009; Fraser, 2004, 2011; Fraser and Naderi, 2008; Fraser *et al.*, 2005, 2010). *Objective Bayesian inference*, which aims at finding non-subjective model based priors, is also part of this effort. Examples of recent breakthroughs related to *reference prior* and model selction are Bayarri *et al.* (2012); Berger (1992); Berger and Sun (2008); Berger *et al.* (2009, 2012). Additional fiducial related works include Wang (2000); Xu and Li (2006); Veronese and Melilli (2014) or Taraldsen and Lindqvist (2013) who show how fiducial distributions naturally arise within a decision theoretical framework; and this list is by no means exhaustive.

Arguably, *Generalized Fiducial Inference* (GFI) has been on the forefront of the modern fiducial revival. It is motivated by the work of Tsui and Weerahandi (1989, 1991) and Weerahandi (1993, 1994, 1995) on *generalized confidence intervals* and the work of Chiang (2001) on the *surrogate variable method* for obtaining confidence intervals for variance components. The main spark came from the realization that there was a connection between these new procedures and fiducial inference. This realization evolved through a series of works Iyer *et al.* (2004); Patterson *et al.* (2004); Hannig *et al.* (2006b); Hannig (2009).

GFI defines a data dependent measure on the parameter space by carefully using an inverse of a deterministic *data generating equation* without the use of Bayes theorem. The resulting *generalized fiducial distribution* (GFD) is a data dependent distribution on the parameter space. GFD can be viewed as a distribution estimator (as opposed to a point or interval estimator) of the unknown parameter of interest. The resulting GFD when used to define approximate confidence sets is often shown in simulations to have very desirable properties; e.g., conservative coverages but shorter expected lengths than competing procedures (E *et al.*, 2008).

The strengths and limitations of the generalized fiducial approach are becoming better understood, see, especially, Hannig (2009, 2013). In particular, the asymptotic exactness of fiducial confidence sets, under fairly general conditions, was established in Hannig (2013); Hannig *et al.* (2006b); Sonderegger and Hannig (2014). Higher order asymptotics of GFI was studied in Majumder and Hannig (2015). GFI has also been extended to prediction problems in Wang *et al.* (2012a). Model selection was introduced into the GFI paradigm in Hannig and Lee (2009). This idea was then further explored in classical setting in Wandler and Hannig (2011) and in the ultra-high-dimensional regression in Lai *et al.* (2015)

GFI has been proven useful in many practical applications. Earlier examples include bioequivalence (McNally *et al.*, 2003; Hannig *et al.*, 2006a), problems of metrology (Hannig *et al.*, 2003, 2007; Wang and Iyer, 2005, 2006a,b; Wang *et al.*, 2012b) and interlaboratory experiments and international key comparison experiments (Iyer *et al.*, 2004). It has also been applied to derive confidence procedures in many important statistical problems, such as variance components (Cisewski and Hannig, 2012; E *et al.*, 2008), maximum mean of a multivariate normal distribution (Wandler and Hannig, 2011), multiple comparisons (Wandler and Hannig, 2012a), extreme value estimation (Wandler and Hannig, 2012b), mixture of normal and Cauchy distributions (Glagovskiy, 2006), wavelet regression (Hannig and Lee, 2009) and logistic regression and binary response models (Liu and Hannig, 2014).

One main goal of this paper is to deliver a concise introduction to GFI. Our intention is to

provide a single location where the various developments of the last decade can be found. As a second goal of this paper, some original work and refined results on GFI are also presented. Specifically, they are Definition 2.1 and Theorems 2.1, 2.3 and 3.1.

The rest of this paper is organized as follows. Starting from Fisher's fiducial argument, Section 2 provides a complete description of GFI, including some new results. The issue of model selection within the GFI framework is discussed in Section 3. Section 4 concerns the use of GFI for discrete and discretized data, and Section 5 offers some practical advice to handle common computational challenges when applying GFI. Lastly, Section 5.1 provides some concluding remarks while technical details are relegated to the appendix. The following website `http://anson.ucdavis.edu/~tcmlee/GFiducial.html` contains computer code for many of the methods in this review.

2 The Switching Principle: Fisher's "Fiducial Argument" Extended

The idea underlying GFI is motivated by our understanding of Fisher's fiducial argument. GFI begins with expressing the relationship between the data, $\boldsymbol{Y}$, and the parameters, $\boldsymbol{\theta}$, as

$$\boldsymbol{Y} = \boldsymbol{G}(\boldsymbol{U}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{G}(\cdot, \cdot)$ is a deterministic function termed the *data generating equation*, and $\boldsymbol{U}$ is the random component of this data generating equation whose distribution is independent of parameters and completely known.

The data $\boldsymbol{Y}$ are assumed to be created by generating a random variable $\boldsymbol{U}$ and plugging it into the data generating equation (1). For example a single observation from $N(\mu, 1)$ distribution can be written as $Y = \mu + U$, where $\boldsymbol{\theta} = \mu$ and U is $N(0, 1)$ random variable.

For simplicity, this subsection only considers the simple case where the data generating equation (1) can be inverted and the inverse $Q_{\boldsymbol{y}}(\boldsymbol{u}) = \boldsymbol{\theta}$ exists for any observed $\boldsymbol{y}$ and for any

arbitrary $\boldsymbol{u}$. Fisher's *Fiducial Argument* leads one to define the fiducial distribution for $\boldsymbol{\theta}$ as the distribution of $Q_{\boldsymbol{y}}(\boldsymbol{U}^\star)$ where $\boldsymbol{U}^\star$ is an independent copy of $\boldsymbol{U}$. Equivalently, a sample from the fiducial distribution of $\boldsymbol{\theta}$ can be obtained by generating $\boldsymbol{U}_i^\star,\ i = 1, \ldots, N$ and using $\boldsymbol{\theta}_i^\star = Q_{\boldsymbol{y}}(\boldsymbol{U}_i^\star)$. Estimates and confidence intervals for $\boldsymbol{\theta}$ can be obtained based on this sample. In the $N(\mu, 1)$ example, $Q_y(u) = y - u$ and the fiducial distribution is therefore the distribution of $y - U^\star \sim N(y, 1)$.

Example 2.1. Consider the mean and sample variance $\boldsymbol{Y} = (\bar{Y}, S^2)$ computed from n independent $N(\mu, \sigma^2)$ random variables, where μ and σ^2 are parameters to be estimated. A natural data generating equation for $\boldsymbol{Y}$ is

$$\bar{Y} = \mu + \sigma U_1 \text{ and } S^2 = \sigma^2 U_2,$$

where U_1, U_2 are independent with $U_1 \sim N(0, n^{-1})$ and $U_2 \sim \text{Gamma}((n-1)/2, (n-1)/2)$.

The inverse $Q_{\boldsymbol{y}}(\boldsymbol{u}) = (\bar{y} - s\, u_1/\sqrt{u_2}, s^2/u_2)$. Consequently, for any observed value $\bar{y}$ and s, and an independent copy of $\boldsymbol{U}$, denoted as $\boldsymbol{U}^\star$, the distribution of $\mu^\star = \bar{y} - s\, U_1^\star/\sqrt{U_2^\star}$ is the marginal fiducial distribution of μ. The equal tailed set of 95% fiducial probability is $(\bar{y} - ts/\sqrt{n}, \bar{y} + ts/\sqrt{n}\ )$ where t is the 0.025 critical value of the t-distribution with $n-1$ degrees of freedom which is the classical 95% confidence interval for μ.

Remark 2.1. We have made a conscious choice to eschew philosophical controversies throughout the development of GFI. However, we find it inevitable to make at least some philosophical comments at this point:

1. The idea behind GFD is very similar to the idea behind the likelihood function: what is the chance of observing my data if any given parameter was true. The added value of GFD is that it provides likelihood function with an appropriate Jacobian obtaining a proper probability distribution on the parameter space, see (4) below.

2. GFD does not presume that the parameter is random. Instead it should be viewed as a distribution estimator (rather than a point or interval estimator) of the fixed true

parameter. To validate this distribution estimator in a specific example we then typically demonstrate good small sample performance by simulation and prove good large sample properties by asymptotic theorems.

3. From a Bayesian point of view, Bayes theorem updates the distribution of $\boldsymbol{U}$ after the data are observed. However, when no prior information is present, changing the distribution of $\boldsymbol{U}$ only by restricting it to the set "there is at least one $\boldsymbol{\theta}$ solving the equation $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{U}, \boldsymbol{\theta})$" seems to us as a reasonable choice (see next section). Arguably, this so called "continuing to believe" assumption has been behind most of the philosophical controversies surrounding fiducial inference in the past.

2.1 A Refined Definition of Generalized Fiducial Distribution

The inverse to equation (1) does not exist for two possible reasons. Either, there is more than one $\boldsymbol{\theta}$ for some value of $\boldsymbol{y}$ and $\boldsymbol{u}$, or there is no $\boldsymbol{\theta}$ satisfying $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})$. The first situation can be dealt with by using the mechanics of Dempster-Shafer calculus (Dempster, 2008). A more practical solution is to select one of the several solutions using a possibly random mechanism. In Section 4 we will review theoretical results that showed that the uncertainty due to multiple solutions has, in many parametric problems, only a second order effect on statistical inference.

For the second situation, Hannig (2009) suggests[1] removing the values of $\boldsymbol{u}$ for which there is no solution from the sample space and then re-normalizing the probabilities; i.e., using the distribution of $\boldsymbol{U}$ conditional on the event $\mathcal{U}_{\boldsymbol{y}} = \{\boldsymbol{u} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}), \text{ for some } \boldsymbol{\theta}\}$. The rationale for this choice is that we know that the observed data $\boldsymbol{y}$ were generated using some

[1] An exception to this suggestion is in cases where the parameter space is in some way constrained. In this case it is often beneficial to extend the parameter space, perform the inversion in the extended space and then project to the boundary of the constrained parameter space. A good example of such a situation is the variance component model where variances are constrained to be greater than or equal to zero (E *et al.*, 2009; Cisewski and Hannig, 2012).

fixed unknown $\boldsymbol{\theta}_0$ and $\boldsymbol{u}_0$; i.e., $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{\theta}_0, \boldsymbol{u}_0)$. The values of $\boldsymbol{u}$ for which $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{.}\,, \boldsymbol{u})$ does not have a solution could not be the true $\boldsymbol{u}_0$ hence only the values of $\boldsymbol{u}$ for which there is a solution should be considered in the definition of the *generalized fiducial distribution*. However, $\mathcal{U}_{\boldsymbol{y}}$, the set of $\boldsymbol{u}$ for which the solution exists, has probability zero for most problems involving absolutely continuous random variables. Conditioning on such a set of probability zero will therefore lead to non-uniqueness due to the Borel paradox (Casella and Berger, 2002, Section 4.9.3).

Hannig (2013) proposes an attractive interpretation of the conditional distribution by limit of discretizations. Here we generalize this approach slightly. Throughout this manuscript $\boldsymbol{U}^\star$ denotes an independent copy of $\boldsymbol{U}$ and $\boldsymbol{\theta}^\star$ denotes a random variable taking values in the parameter space $\boldsymbol{\Theta}$.

To define GFD we need to interpret the ill-defined conditional distribution of $\boldsymbol{U}^\star \mid \boldsymbol{U}^\star \in \mathcal{U}_{\boldsymbol{y}}$. To do that we "fatten up" the manifold $\mathcal{U}_{\boldsymbol{y}}$ by ϵ so that the enlarged set $\mathcal{U}_{\boldsymbol{y},\epsilon} = \{\boldsymbol{u} : \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\| \leq \epsilon, \text{ for some } \boldsymbol{\theta}\}$ has positive probability and the conditional distribution of $\boldsymbol{U}^\star \mid \boldsymbol{U}^\star \in \mathcal{U}_{\boldsymbol{y},\epsilon}$ is well defined. Finally the fattening needs to be done in a consistent way so that the limit of conditional distributions as $\epsilon \to 0$ is well defined. This leads to the following definition:

Definition 2.1. A probability measure on the parameter space $\boldsymbol{\Theta}$ is called a *generalized fiducial distribution* (GFD) if it can be obtained as a weak limit

$$\lim_{\epsilon \to 0} \left[\arg\min_{\boldsymbol{\theta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\| \;\middle|\; \min_{\boldsymbol{\theta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\| \leq \epsilon \right]. \tag{2}$$

If there are multiple minimizers $\arg\min_{\boldsymbol{\theta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\|$, one selects one of them (potentially at random). Notice that the conditioning in (2) is modifying the distribution of $\boldsymbol{U}^\star$ in order to only consider values for which an approximate inverse to $\boldsymbol{G}$ exists.

Remark 2.2. Definition 2.1 illuminates the relationship between GFD and Approximate Bayesian Computations (ABC) (Beaumont *et al.*, 2002). In an idealized ABC, one generates first an

observation $\boldsymbol{\theta}^*$ from the prior, then generates a new sample using a data generating equation $\boldsymbol{y}^\star = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)$ and compares the generated data with the observed data $\boldsymbol{y}$. If the observed and generated data sets are close (e.g., $\|\boldsymbol{y} - \boldsymbol{y}^\star\| \leq \epsilon$), the generated $\boldsymbol{\theta}^\star$ is accepted, otherwise it is rejected and the procedure is repeated. If the measure of closeness is a norm, it is easy to see that when $\epsilon \to 0$ the weak limit of the ABC distribution is the posterior distribution.

On the other hand, when defining GFD one generates $\boldsymbol{U}^\star$, finds a best fitting $\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\|$, computes $\boldsymbol{y}^\star = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)$, again accepts $\boldsymbol{\theta}^\star$ if $\|\boldsymbol{y} - \boldsymbol{y}^\star\| \leq \epsilon$ and rejects otherwise.

In either approach an artificial data set $\boldsymbol{y}^\star = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)$ is generated and compared to the observed data. The main difference is that the Bayes posterior simulates the parameter $\boldsymbol{\theta}^\star$ from the prior while GFD uses the best fitting parameter.

Remark 2.3. The GFD defined in (2) is not unique as it depends on both the data generating equation (1), the norm used in (2) and the minimizer $\boldsymbol{\theta}^\star$ chosen. Let $\boldsymbol{U}^\star$ be an independent copy of $\boldsymbol{U}$ and let for any measurable set A, $V[A]$ be a rule selecting a possibly random element of the closure of the set $\bar{A}$. When the probability $P(\exists \boldsymbol{\theta}^\star,\ \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)) > 0$ then the limit (2) is the conditional distribution

$$V[\{\boldsymbol{\theta}^\star : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\}] \mid \{\exists \boldsymbol{\theta}^\star,\ \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\}.$$

This is an older definition of GFD that can be found in Hannig (2009, 2013).

The next subsection offers a useful computational formula for evaluating (2).

2.2 A User Friendly Formula for Generalized Fiducial Distribution

While Definition (2) for GFD is conceptually appealing and very general, it is not immediately clear how to compute the limit in many practical situations. In a less general setup using the l^∞ norm, Hannig (2013) derives a closed form of the limit in (2) applicable to many practical

situations. Here we provide a generalization of this result, which is applicable in most situations where the data follows a continuous distribution.

Assume that the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ is p-dimensional, the data $\boldsymbol{x} \in \mathbb{R}^n$ are n dimensional. The following theorem provides a useful computational formula.

Theorem 2.1. *Supppose Assumptions A.1 to A.3 stated in Appendix A. Then the limiting distribution in (2) has a density*

$$r(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y},\boldsymbol{\theta})J(\boldsymbol{y},\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y},\boldsymbol{\theta}')J(\boldsymbol{y},\boldsymbol{\theta}')\,d\boldsymbol{\theta}'}, \tag{3}$$

where $f(\boldsymbol{y},\boldsymbol{\theta})$ is the likelihood and the function

$$J(\boldsymbol{y},\boldsymbol{\theta}) = D\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\Big|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right). \tag{4}$$

If (i) $n = p$ then $D(A) = |\det A|$. Otherwise the function $D(A)$ depends on the norm used; (ii) the l_∞ norm gives $D(A) = \sum_{\mathbf{i}=(i_1,\ldots,i_p)} |\det(A)_{\mathbf{i}}|$; (iii) under an additional Assumption A.4 the l_2 norm gives $D(A) = (\det A^\top A)^{1/2}$.

In (ii) the sum spans over $\binom{n}{p}$ of p-tuples of indexes $\boldsymbol{i} = (1 \le i_1 < \cdots < i_p \le n)$. For any $n \times p$ matrix A, the sub-matrix $(A)_{\boldsymbol{i}}$ is the $p \times p$ matrix containing the rows $\boldsymbol{i} = (i_1, \ldots, i_p)$ of A. Based on our current experience we recommend using (ii) in practice.

There is a slight abuse of notation in (3) as $r(\boldsymbol{\theta}|\boldsymbol{y})$ is not a conditional density in the usual sense. Instead, we are using this notation to remind the reader that the fiducial density depends on the fixed observed data.

Cases (i) and (ii) are a simple consequence of results in Hannig (2013). The formula in (iii) was independently proposed in Fraser *et al.* (2010) based on arguments related to tangent exponential families without being recognized as a fiducial distribution. The proof is in Appendix A. The ease of use of (4) will be demonstrated on several examples in the next subsection. The rest of this subsection discusses the effects of various transformations.

Remark 2.4. *Just like posterior computed using Jeffreys prior, GFD is invariant under smooth re-parametrizations.*

This assertion has been shown for smooth transformation by chain rule in Hannig (2013). However this property is general and follows directly from (2), since for an appropriate selection of minimizers and any one-to-one function $\boldsymbol{\theta} = \phi(\boldsymbol{\eta})$

$$\phi\left(\arg\min_{\boldsymbol{\eta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \phi(\boldsymbol{\eta}^\star))\|\right) = \arg\min_{\boldsymbol{\theta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\|.$$

Remark 2.5. *GFD could change with transformations of the data generating equation.*

Assume that the observed data set has been transformed with a one-to-one smooth transformation $\boldsymbol{Z} = \boldsymbol{T}(\boldsymbol{Y})$. Using the chain rule we see that the GFD based on this new data generating equation and with observed data $\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{y})$ is the density (3) with the Jacobian function (4) simplified to

$$J_{\boldsymbol{T}}(\boldsymbol{z}, \boldsymbol{\theta}) = D\left(\frac{\boldsymbol{d}}{\boldsymbol{dy}}\boldsymbol{T}(\boldsymbol{y}) \cdot \frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\Big|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right). \tag{5}$$

Notice that for simplicity we write $\boldsymbol{y}$ instead of $T^{-1}(\boldsymbol{z})$ in (5).

For completeness we recall the well known fact that the likelihood based on $\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{y})$ satisfies

$$f_{\boldsymbol{T}}(\boldsymbol{z}|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})\left|\det\left(\frac{\boldsymbol{dT}}{\boldsymbol{dy}}\right)\right|^{-1}. \tag{6}$$

The second term in the right-hand-side of (6) is a constant and does not affect the GFD with the exception of model selection considerations in Section 3.

As can be seen from the above calculation GFD will usually change with transformation of the data. An important exception is when the number of observations and number of parameters are equal; i.e., $n = p$. Indeed, by careful evaluation of (4), (5) and (6) we see that for $\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{y})$ we have $J(\boldsymbol{y}, \boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = J_{\boldsymbol{T}}(\boldsymbol{z}, \boldsymbol{\theta}) f_{\boldsymbol{T}}(\boldsymbol{z}|\boldsymbol{\theta})$ and the GFD is unchanged.

Example 2.2. Consider the following important transformation. Let $\boldsymbol{Z} = (\boldsymbol{S}, \boldsymbol{A})^\top$, where $\boldsymbol{S}$ is a p-dimensional sufficient statistic and $\boldsymbol{A}$ is an ancillary statistic. Let $\boldsymbol{s} = \boldsymbol{S}(\boldsymbol{y})$ and $\boldsymbol{a} = \boldsymbol{A}(\boldsymbol{y})$

be the observed values of the sufficient and ancillary statistics respectively. Since $\boldsymbol{dA}/\boldsymbol{d\Theta} = 0$, the function D in (5) is the absolute value of the determinant of the $p \times p$ non-zero sub-matrix:

$$J(\boldsymbol{z}, \boldsymbol{\theta}) = \left| \det \left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}} \boldsymbol{S}(\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})) \Big|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})} \right) \right|. \tag{7}$$

Next, denote the solution of the equation $\boldsymbol{s} = \boldsymbol{S}(\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}))$ by $Q_{\boldsymbol{s}}(\boldsymbol{u}) = \boldsymbol{\theta}$. A straightforward calculation shows that the fiducial density (3) with (7) is the conditional distribution of $Q_{\boldsymbol{s}}(\boldsymbol{U}^\star) \mid \boldsymbol{A}(\boldsymbol{U}^\star) = \boldsymbol{a}$, the GFD based on sufficient statistic conditional on ancillary, c.f. Birnbaum (1962); Iyer and Patterson (2002).

2.3 Simple Examples

In this section we will consider two examples, linear regression and uniform distribution. In the first case the GFD is the same as Bayes posterior with respect to the independence Jeffreys prior while in the second the GFD is not a Bayes posterior with respect to any prior (that is not data dependent).

Example 2.3 (Linear Regression). Express linear regression using the data generating equation

$$\boldsymbol{Y} = G(\boldsymbol{U}, \boldsymbol{\theta}) = \boldsymbol{X}\beta + \sigma\boldsymbol{U},$$

where $\boldsymbol{Y}$ is the dependent variables, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\theta} = (\beta, \sigma)$ are the unknown parameters and $\boldsymbol{U}$ is a random vector with known density $f(\boldsymbol{u})$ independent of any parameters.

In order to compute GFD simply notice that $\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{U}, \boldsymbol{\theta}) = (\boldsymbol{X}, \boldsymbol{U}), \quad \boldsymbol{U} = \sigma^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta)$. From here the Jacobian in (4) using the l_∞ norm simplifies to

$$J_\infty(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-1} \sum_{\substack{\boldsymbol{i}=(i_1,\ldots,i_p) \\ 1 \le i_1 < \cdots < i_p \le n}} \left| \det \left(\boldsymbol{X}, \boldsymbol{Y} \right)_{\boldsymbol{i}} \right|$$

and the density of GFD is

$$r(\beta, \sigma | \boldsymbol{y}) \propto \sigma^{-n-1} f(\sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)).$$

This coincides with the Bayesian solution using the independence Jeffreys prior (Berger, 2011).

The J function has a more compact form when using the l_2 norm. In particular by Cauchy-Binet formula we see that $\det\big((\boldsymbol{X}, \boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{X}, \boldsymbol{y} - \boldsymbol{X}\beta)\big)$ is invariant in β. By selecting $\beta = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ we immediately obtain

$$J_2(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-1}|\det(\boldsymbol{X}^\top\boldsymbol{X})|^{\frac{1}{2}}\mathrm{RSS}^{\frac{1}{2}},$$

where RSS is the residual sum of squares. As the two Jacobian functions differ only by a constant, the GFD is unchanged.

As a special case the GFD for the location-scale model $\boldsymbol{X} = \boldsymbol{1}$, the l_∞ Jacobian is $J_\infty(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-1}\sum_{i<j}|Y_i - Y_j|$ while the l_2 Jacobian becomes $J_2(\boldsymbol{y}, \boldsymbol{\theta}) = \sigma^{-1}n\hat{\sigma}_n$, where $\hat{\sigma}_n$ is the maximum likelihood estimator of σ.

Example 2.4 (Uniform $U\{a(\theta) - b(\theta), a(\theta) + b(\theta)\}$). As a second example we will study a very irregular model. The reference prior for this model is complicated and has been obtained as Theorem 8 in Berger *et al.* (2009).

Express the observed data using the following data generating equation

$$Y_i = a(\theta) + b(\theta)U_i, \quad U_i \text{ i.i.d. } U(-1, 1).$$

Simple computations give $\frac{d}{d\theta}\boldsymbol{G}(\boldsymbol{u}, \theta) = a'(\theta) + b'(\theta)\boldsymbol{U}$ with $\boldsymbol{U} = b^{-1}(\theta)(\boldsymbol{Y} - a(\theta))$. If $a'(\theta) > |b'(\theta)|$, (4) simplifies to

$$J(\boldsymbol{y}, \theta) = \sum_{i=1}^{n}\big|a'(\theta) + \{\log b(\theta)\}'\{y_i - a(\theta)\}\big| = n[a'(\theta) - a(\theta)\{\log b(\theta)\}' + \bar{y}_n\{\log b(\theta)\}'].$$

We used $a'(\theta) > |b'(\theta)|$ only to show that the terms inside the absolute values below are all positive. However we remark that under this assumption both $a(\theta) - b(\theta)$ and $a(\theta) + b(\theta)$ are strictly increasing, continuous functions of θ.

With the above, the GFD is then

$$r(\theta|\boldsymbol{y}) \propto \frac{a'(\theta) - a(\theta)\{\log b(\theta)\}' + \bar{y}_n\{\log b(\theta)\}'}{b(\theta)^n} I_{\{a(\theta)-b(\theta)<y_{(1)} \ \&\ a(\theta)+b(\theta)>y_{(n)}\}}. \tag{8}$$

To demonstrate how the Jacobian changes with transformation of the data, consider a transformation of the data $\boldsymbol{Z} = \{Y_{(1)}, Y_{(n)}, (\boldsymbol{Y} - Y_{(1)})/(Y_{(n)} - Y_{(1)})\}^\top$ inspired by Example 2.2. There are only two non-zero terms in (5) and consequently

$$J_{\boldsymbol{Z}}(\boldsymbol{y}, \theta) = 2\left[a'(\theta) - a(\theta)\{\log b(\theta)\}' + \frac{y_{(1)} + y_{(n)}}{2}\{\log b(\theta)\}'\right].$$

We performed a small scale simulation study for $U(\theta, \theta^2)$; $a(\theta) = \theta, b(\theta) = \theta^2 - \theta$. For all the combinations of $n = 1, 2, 3, 4, 5, 10, 20, 100$ and $\theta = 1.01, 1.5, 2, 10, 50, 250$ we analyzed 8000 independent datasets. The results reveal that even though the coverage of the 95% one sided and two sided fiducial credible sets is guaranteed only asymptotically (see below) the coverage appears indistinguishable from the nominal value even for n as small as 1. This is true for both fiducial proposals. Also both fiducial distributions concentrate closer to the true value (as measured by mean absolute deviations) than Bayesian posteriors with flat and reference priors (Berger *et al.*, 2009).

2.4 Theoretical Results

This section discusses asymptotic properties for GFI. We hope that the material included here will be useful for the study of GFD in future practical problems.

First we present a Bernstein-von Mises theorem for GFD, which provides theoretical guarantees of asymptotic normality and asymptotic efficiency. It also guarantees in conjunction with Theorem 2.3 below that appropriate sets of fiducial probability $1 - \alpha$ are indeed approximate $1 - \alpha$ confidence sets. An early version of this theorem can be found in Hannig (2009). Here we will state a more general results due to Sonderegger and Hannig (2014).

Assume that we are given a random sample of independent observations $Y_1, \ldots, Y_n$ with data generating equation $Y_i = \boldsymbol{G}(\boldsymbol{\theta}, U_i)$, with U_i i.i.d. $U(0, 1)$. This data generating equation leads to the Jacobian function (4) that is a U-statistic. This realization makes the GFD amenable to theoretical study.

Asymptotic normality of statistical estimators usually relies on a set of technical assumptions and GFD is no exception. In order to succinctly state the theorem we denote the rescaled density of GFD by $r^*(\boldsymbol{s}|\boldsymbol{y}) = n^{-1/2} r(n^{-1/2}\boldsymbol{s} + \hat{\boldsymbol{\theta}}|\boldsymbol{y})$, where $\hat{\boldsymbol{\theta}}$ is the consistent MLE.

Theorem 2.2 (Sonderegger and Hannig, 2014)**.** *Under Assumptions B.1 to B.4 in Appendix B*

$$\int_{\mathbb{R}^p} \left| r^*(\boldsymbol{s}|\boldsymbol{y}) - \frac{\sqrt{det\,|I(\boldsymbol{\theta}_0)|}}{\sqrt{2\pi}} e^{-\boldsymbol{s}^T I(\boldsymbol{\theta}_0)\mathbf{s}/2} \right| d\boldsymbol{s} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0.$$

One application of GFD is to take sets of $1-\alpha$ fiducial probability and use them as approximate confidence intervals. Next we state conditions under which this is valid.

Assumption 2.1. Let us consider a sequence of data sets $\boldsymbol{Y}_n$ generated using fixed parameters $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}_n$ with corresponding data dependent measures[2] on the parameter space $R_{n,\boldsymbol{Y}_n}$. We will assume that these converge to a limiting fiducial model in the following way:

1. There is a sequence of measurable functions t_n of $\boldsymbol{Y}_n$ so that $t_n(\boldsymbol{Y}_n)$ converges in distribution to some random variable $\boldsymbol{T}$.

2. (a) The $\boldsymbol{T}$ from Part 1 can be decomposed into $\boldsymbol{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2)$ and there is a limiting data generating equation $\boldsymbol{T}_1 = \boldsymbol{H}_1(\boldsymbol{V}_1, \boldsymbol{\xi}), \boldsymbol{T}_2 = \boldsymbol{H}_2(\boldsymbol{V}_2)$, where $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2)$ has a fully known distribution independent of the parameter $\boldsymbol{\xi} \in \boldsymbol{\Xi}$. The distribution of $\boldsymbol{T}$ is obtained from the limiting data generating equation using $\boldsymbol{\xi}_0$.

   (b) The equation $\boldsymbol{H}_1$ is one-to-one if viewed as a function (possibly implicit) for any combination of $\boldsymbol{\xi}$, $\boldsymbol{v}_1$, $\boldsymbol{t}_1$, where one is held fixed, one taken as a dependent and one taken as an independent variable. The equation $\boldsymbol{H}_2$ is one to one. Consequently, the limiting GFD defined by (2) is the conditional distribution $Q_{\boldsymbol{t}_1}(\boldsymbol{V}_1^\star) \mid \boldsymbol{H}_2(\boldsymbol{V}_2^\star) = \boldsymbol{t}_2$, where $Q_{\boldsymbol{t}_1}(\boldsymbol{v}_1) = \boldsymbol{\xi}$ is the solution of $\boldsymbol{t}_1 = \boldsymbol{H}_1(\boldsymbol{v}_1, \boldsymbol{\xi})$. Denote this conditional measure by $R_{\boldsymbol{t}}$.

[2]Such data dependent measures can be for example GFDs, Bayes posteriors or a confidence distributions.

(c) For any open set $C \subset \boldsymbol{\Xi}$ and limiting data $\boldsymbol{t}$ the limiting fiducial probability of the boundary $R_{\boldsymbol{t}}(\partial C) = 0$.

3. There are homeomorphic injective mappings Ξ_n from $\boldsymbol{\Theta}_n$ into $\boldsymbol{\Xi}$ so that

   (a) $\Xi_n(\boldsymbol{\theta}_{n,0}) = \boldsymbol{\xi}_0$;

   (b) For any sequence of data $t_n(\boldsymbol{y}_n) \to \boldsymbol{t}$ the transformed fiducial distribution measures converge weakly $R_{n,\boldsymbol{y}_n}\Xi_n^{-1} \xrightarrow{\mathcal{W}} R_{\boldsymbol{t}}$.

Theorem 2.3. *Suppose Assumption 2.1 holds. Fix a desired coverage* $0 < \alpha < 1$. *For any observed data* $\boldsymbol{y}_n$ *select an open set* $C_n(\boldsymbol{y}_n)$ *satisfying: (i)* $R_{n,\boldsymbol{y}_n}(C_n(\boldsymbol{y}_n)) = \alpha$; *(ii)* $t_n(\boldsymbol{y}_n) \to \boldsymbol{t}$ *implies* $\Xi_n(C_n(\boldsymbol{y}_n)) \to C(\boldsymbol{t})$; *(iii) the set* $\mathcal{V}_{\boldsymbol{t}_2} = \{(\boldsymbol{v}_1, \boldsymbol{v}_2) : Q_{\boldsymbol{t}_1}(\boldsymbol{v}_1) \in C(\boldsymbol{t})$ *and* $\boldsymbol{t}_2 = \boldsymbol{H}_2(\boldsymbol{v}_2)\}$ *is invariant in* $\boldsymbol{t}_1$. *Then the sets* $C_n(\boldsymbol{y}_n)$ *are* α *asymptotic confidence sets.*

The theorem provides a condition on how various sets of a fixed fiducial probability need to be linked together across different observed data sets in order to make up a valid confidence set. To understand the key Condition (iii) notice that it assumes the sets $C(\boldsymbol{t})$ are obtained by de-pivoting a common set $\mathcal{V}_{\boldsymbol{t}_2}$. In particular if the limiting data generating equation $\boldsymbol{T}_1 = \boldsymbol{H}_1(\boldsymbol{V}_1, \boldsymbol{\xi})$ has group structure, Condition (iii) is equivalent to assuming the sets $C(\boldsymbol{t})$ are group invariant in $\boldsymbol{t}_1$. The conditions on the limiting data generating equation were partially inspired by results for Inferential Models of Martin and Liu (2015a). The proof of Theorem 2.3 is in Appendix C. Also, this corollary follows immediately:

Corollary 2.1. *Any model that satisfies the assumptions of Theorem 2.2 satisfies Assumption 2.1. In particular, for any fixed interior point* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}^0$ *the limiting data generating equation* $\boldsymbol{T} = \boldsymbol{\xi} + \boldsymbol{V}$ *where the random vector* $\boldsymbol{V} \sim N(0, I(\boldsymbol{\theta}_0)^{-1})$. *The transformations are* $t_n(\boldsymbol{y}_n) = n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, $\Xi_n(\boldsymbol{\theta}) = n^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ *and* $\boldsymbol{\xi}_0 = 0$. *Any collection of sets* $C_n(\boldsymbol{y}_n)$ *that in the limit becomes location invariant will form asymptotic confidence intervals.*

Most of the theoretical results for GFI in the literature were derived in regular statistical

problems and are covered by Corollary 2.1. Notice that in the regular case the limiting data generating equation has no ancillary part. The next example shows that the ancillary part in Theorem 2.3 is needed in some non-regular cases.

Example 2.5 (Example 2.4 continued). Recall that $Y_i = a(\theta) + b(\theta)U_i,\ i = 1, \ldots, n$, where U_i are i.i.d. $U(-1, 1)$. We assume that $a'(\theta) > |b'(\theta)|$ for $\theta \in \boldsymbol{\Theta}$ so that the GFD R_{n,y_n} has a density given by (8). Fix an interior point $\theta_0 \in \boldsymbol{\Theta}^0$. In order to verify conditions of Theorem 2.3 we need to define the limiting data generating equation, and the transformations t_n and Ξ_n. We start with the limiting data generating process:

$$T_1 = \xi + V_1, \quad T_2 = V_2,$$

where $V_1 = (E_1 - E_2)/2$, $V_2 = (E_1 + E_2)/2$ with E_1, E_2 are independent, $E_1 \sim \text{Exp}[\{a'(\theta_0) - b'(\theta_0)\}/\{2b(\theta_0)\}]$ and $E_2 \sim \text{Exp}[\{a'(\theta_0) + b'(\theta_0)\}/\{2b(\theta_0)\}]$. The density of the limiting GFD is therefore proportional to

$$r(\xi|\boldsymbol{t}) \propto e^{-\xi\,\{\log b(\theta_0)\}'} I_{(T_1-T_2, T_1+T_2)}(\xi).$$

The fact that Assumption 2.1, Part 2 is satisfied follows immediately.

Next, define the transformations

$$t_n(\boldsymbol{y}) = n \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix} \cdot \begin{pmatrix} \frac{y_{(1)} - (a(\theta_0) - b(\theta_0))}{a'(\theta_0) - b'(\theta_0)} \\ \frac{a(\theta_0) + b(\theta_0) - y_{(n)}}{a'(\theta_0) + b'(\theta_0)} \end{pmatrix}, \quad \Xi_n(\theta) = n(\theta - \theta_0).$$

Simple calculations show that Assumption 2.1, Part 1 and 3 are satisfied with $\boldsymbol{\xi}_0 = 0$.

Finally, notice that any collection of sets of fiducial probability α that in the limit becomes location invariant in t_1 (such as one sided or equal tailed intervals) are asymptotic α confidence intervals.

2.5 Practical Use of GFI

From a practical point of view GFI is used in a way similar to the use of a posterior computed using a default (objective) prior, such as probability matching, reference or flat prior. The

main technical difference is that the objective prior is replaced by a data dependent Jacobian (4). This data dependence can in some examples lead to the existence of second order matching GFD even when only first order matching is available with the non-data dependent priors (Majumder and Hannig, 2015). Some argued (Welch and Peers, 1963; Martin and Walker, 2014) that data dependent priors are essential in achieving superior frequentist properties in complex statistical problems.

First, we suggest using a set of fiducial probability $1-\alpha$ and of a good shape (such as one sided or equal tailed) as an approximate $1-\alpha$ confidence interval, see Theorem 2.3. Next, the mean or median of the GFD can be used for point estimation.

GFDs can also be used for predicting future observations. This is done by plugging in a random variable having the GFD (2) for the parameter into the data generating equation for the new observations. This approach produces a predictive distribution that accommodates in a natural way both the uncertainty in the parameter estimation and the randomness of the future data. More details are in Wang *et al.* (2012a).

GFDs are rarely available in closed form. Therefore we often need to use an MCMC method such as a Metropolis-Hastings or Gibbs sampler to obtain a sample from the GFD. While the basic issues facing implementation of the MCMC procedures are similar for both Bayesian and generalized fiducial problems, there are specific challenges related to generalized fiducial procedures. We discuss some computational issues in Section 5.

3 Model Selection in GFI

Hannig and Lee (2009) introduce model selection into the GFI paradigm in the context of wavelet regression. The presentation here is re-expressed using definition (2). There are two main ingredients needed for an effective fiducial model selection. The first is to include the model as one of the parameters and the second is to include penalization in the data generating

equation.

Consider a finite collection of models $\mathcal{M}$. The data generating equation is

$$\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\theta}_M, \boldsymbol{U}), \qquad M \in \mathcal{M},\ \boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M, \tag{9}$$

where $\boldsymbol{Y}$ is the observations, M is the model considered, $\boldsymbol{\theta}_M$ are the parameters associated with model M, and $\boldsymbol{U}$ is a random vector of with fully known distribution independent of any parameters.

Denote the number of parameters in the model M by $|M|$. Similarly to MLE, an important issue needing to be solved is that GFI tends to favor models with more parameters over ones with fewer parameters. Therefore an outside penalty accounting for our preference toward parsimony needs to be incorporated in the model. See Appendix D for more details.

In Hannig and Lee (2009) a novel way of adding a penalty into the GFI framework is proposed. In particular, for each model M they propose augmenting the data generating equation (9) by

$$0 = P_k, \quad k = 1, \ldots, \min(|M|, n), \tag{10}$$

where P_k are i.i.d. continuous random variables with $f_P(0) = q$ independent of $\boldsymbol{U}$, and q is a constant determined by the penalty. (Based on ideas from the minimum description length principle Hannig and Lee (2009) recommend using $q = n^{-1/2}$ as the default penalty.) Notice that the number of additional equations is the same as the number of unknown parameters in the model.

For the augmented data generating equation we have the following theorem. This theorem has never been published before but it does implicitly appear in Hannig and Lee (2009); Lai *et al.* (2015). For completeness we provide a proof in Appendix D.

Theorem 3.1. *Let us suppose the identifiability Assumption D.1 in Appendix D holds and that each of the models satisfy assumptions of Theorem 2.1 (in particular $|M| \leq n$). Then the*

marginal generalized fiducial probability of model M is

$$r(M|\boldsymbol{y}) = \frac{q^{|M|}\int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{y},\boldsymbol{\theta}_M)J_M(\boldsymbol{y},\boldsymbol{\theta}_M)\,d\boldsymbol{\theta}_M}{\sum_{M'\in\mathcal{M}} q^{|M'|}\int_{\boldsymbol{\Theta}_{M'}} f_{M'}(\boldsymbol{y},\boldsymbol{\theta}_{M'})J_{M'}(\boldsymbol{y},\boldsymbol{\theta}_{M'})\,d\boldsymbol{\theta}_{M'}}, \tag{11}$$

where $f_M(\boldsymbol{y},\boldsymbol{\theta}_M)$ is the likelihood and $J_M(\boldsymbol{y},\boldsymbol{\theta}_M)$ is the Jacobian function computed using (4) for each fixed model M.

Remark 3.1. The quantity $r(M|\boldsymbol{y})$ can be used for inference in the usual way. For example, *fiducial factor*: the ratio $r(M_1|\boldsymbol{y})/r(M_2|\boldsymbol{y})$, can be used in the same way as a Bayes factor. As discussed in Berger and Pericchi (2001) one of the issues with the use of improper priors in Bayesian model selection is the presence of arbitrary scaling constant. While this is not a problem when a single model is considered, because the arbitrary constant cancels, it becomes a problem for model selection. An advantage of GFD is that the Jacobian function (4) comes with a scaling constant attached to it. In fact, the fiducial factors are closely related to the *intrinsic factors* of Berger and Pericchi (1996, 2001). This can be seen from the fact that for the minimal training sample ($n = |M|$), we usually have $\int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{y},\boldsymbol{\theta}_M)J_M(\boldsymbol{y},\boldsymbol{\theta}_M)\,d\boldsymbol{\theta}_M = 1$.

Similarly, the quantity $r(M|\boldsymbol{y})$ can also be used for fiducial model averaging much akin to the Bayesian model averaging (Hoeting *et al.*, 1999).

We illustrate the use of this model selection on two examples, wavelet regression (Hannig and Lee, 2009) and ultra-high dimensional regression (Lai *et al.*, 2015).

3.1 Wavelet Regression

Suppose n observed equispaced data points $\{x_i\}_{i=1}^n$ satisfy the following model:

$$X_i = g_i + \epsilon_i,$$

where $\boldsymbol{g} = (g_1,\ldots,g_n)^\top$ is the true unknown regression function and ϵ_i's are independent standard normal random variables with mean 0 and variance σ^2, and $n = 2^{J+1}$ is an integer power of 2.

Most wavelet regression methods consist of three steps. The first step is to apply a forward wavelet transform to the data $\boldsymbol{y}$ and obtain the empirical wavelet coefficients $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$. Here $\boldsymbol{H}$ is the discrete wavelet transform matrix. The second step is to apply a shrinkage operation to $\boldsymbol{y}$ to obtain an estimate $\hat{\boldsymbol{d}}$ for the true wavelet coefficients $\boldsymbol{d} = \boldsymbol{H}\boldsymbol{g}$. Lastly, the regression estimate $\hat{\boldsymbol{g}} = (\hat{g}_1, \ldots, \hat{g}_n)^\top$ for $\boldsymbol{g}$ is computed via the inverse discrete wavelet transform: $\hat{\boldsymbol{g}} = \boldsymbol{H}^\top \hat{\boldsymbol{d}}$. The second step of wavelet shrinkage is important because it is the step where statistical estimation is performed. Hannig and Lee (2009) use GFI to perform the second step. Apparently this is the first published work where Fisher's fiducial idea is applied to a nonparametric problem.

Due to the orthonormality of the discrete wavelet transform matrix $\boldsymbol{H}$, a model for the empirical wavelet coefficients is $\boldsymbol{Y} = \boldsymbol{d} + \sigma\boldsymbol{U}$ with $\boldsymbol{U}$ being a n-dimensional vector of independent $N(0,1)$ random variables. The assumption of sparsity implies that many of the entries in the vector $\boldsymbol{d}$ are zero. This allows us to cast this as a model selection problem, where the model M is the list of non-zero entries. The data generating equation (9) becomes

$$Y_k = \begin{cases} d_k + \sigma U_k, & k \in M, \\ \sigma U_k, & k \in M^{\complement}. \end{cases}$$

Notice that $\boldsymbol{\theta}_M = \{\sigma^2, d_k \ k \in M\}$. As discussed above we augment the data generating equations by (10) with $q = n^{-1/2}$.

It follows from Theorem 3.1 that the GFD has generalized density proportional to

$$r(\sigma^2, \boldsymbol{d}, M) \propto (\sigma^{-2})^{\frac{n}{2}+1} \frac{\sum_{j\in M^{\complement}} |y_j|}{n - |M|} \times \exp\left[-\frac{|M| \log n}{2} - \frac{\left\{\sum_{k\in M}(d_k - y_k)^2 + \sum_{i\in M^{\complement}} y_i^2\right\}}{2\sigma^2} \right] \prod_{i\in M^{\complement}} \delta_0(d_i), \quad (12)$$

where $\delta_0(s)$ is the Dirac function; i.e., $\int_A \delta_0(s)\, ds = 1$ if $0 \in A$ and 0 otherwise. The term $1/(n - |M|)$ is an additional normalization term introduced to account for the number of the elements in the sum above it.

The normalizing constant in (12) cannot be computed in a closed form so a sample from

$r(\sigma^2, \boldsymbol{d}, I)$ will have to be simulated using MCMC techniques. Note that the GFD is defined in the wavelet domain. Hannig and Lee (2009) use the inverse wavelet transform to define a GFD on the function domain.

Additionally Hannig and Lee (2009) also assume that M satisfies a tree condition (Lee, 2002). This condition states that if a coefficient is thresholded, all its descendants have to be thresholded too; the exact formulation is in Hannig and Lee (2009). This constraint greatly reduces the search space and allows for both efficient calculations and clean theoretical results. In the paper they report a simulation study showing small sample performance superior to the alternative methods considered and prove an asymptotic theorem guaranteeing asymptotic consistency of the fiducial model selection.

3.2 Ultrahigh Dimensional Regression

Lai *et al.* (2015) extend the ideas of fiducial model selection to the ultra-high dimensional regression setting. The most natural data generating equation for this model is

$$\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\beta}_M, \sigma^2, \boldsymbol{Z}) = \boldsymbol{X}_M \boldsymbol{\beta}_M + \sigma \boldsymbol{Z},$$

where $\boldsymbol{Y}$ represents the observations, M is the model considered (collection of parameters that are non-zero), $\boldsymbol{X}_M$ is the design matrix for model M, $\boldsymbol{\beta}_M \in \mathbb{R}^{|M|}$ and $\sigma > 0$ are parameters, and $\boldsymbol{Z}$ is a vector of i.i.d. standard normal random variables. For computational expediency they suggest using a sufficient-ancillary transformation which yields the same Jacobian function as the l_2 Jacobian discussed in Section 2.3. The Jacobian function used is

$$J_M(\boldsymbol{y}, \boldsymbol{\theta}_M) = \sigma^{-1} |\det(\boldsymbol{X}'_M \boldsymbol{X}_M)|^{\frac{1}{2}} \mathrm{RSS}_M^{\frac{1}{2}}.$$

The standard MDL penalty $n^{-|M|/2}$ was not designed to handle ultrahigh dimensional problems. Inspired by the EBIC penalty of Chen and Chen (2008), Lai *et al.* (2015) propose extending the penalty by modifying (10) to

$$0 = B_{|M|}, \quad 0 = P_k, \ k = 1, \ldots, |M|,$$

where $|M|$ is the dimension of M, B_m is a Bernoulli$(1 - r_m)$ random variable that penalizes for the number of models that have the same size m; and P_i are i.i.d. continuous random variables with $f_P(0) = q$ independent of B_m that penalize for the size of models. Following the recommendation of Hannig and Lee (2009) we select $q = n^{-1/2}$. Additionally we select $r_m = \binom{p}{m}^{-\gamma}$, where p is the number of parameters in the full model. The second choice is to penalize for the fact that there is a large number of models that all have the same size. The most natural choice is $\gamma = 1$ for which r_m is the probability of randomly selecting a model M from all models of size m. However, in order to match the EBIC penalty of Chen and Chen (2008) we allow for other choices of γ.

We assume that for any size m, the residual vectors $\{\boldsymbol{I} - \boldsymbol{X}_M(\boldsymbol{X}_M^\top \boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^\top\}\boldsymbol{y}/\mathrm{RSS}_M$ are distinct for all the models $M \in \mathcal{M}'$ of size m, so that the identifiability Assumption D.1 is satisfied. Theorem 3.1 implies

$$r(M|\boldsymbol{y}) \propto R_\gamma(M) = \Gamma\left(\frac{n - |M|}{2}\right)(\pi \mathrm{RSS}_M)^{-\frac{n-|M|-1}{2}} n^{-\frac{|M|+1}{2}} \binom{p}{|M|}^{-\gamma}. \tag{13}$$

Similarly to the tree constraint of the previous subsection, Lai *et al.* (2015) additionally reduce the number of models by constructing a class of candidate models, denoted as $\mathcal{M}'$. This $\mathcal{M}'$ should satisfy the following two properties: the number of models in $\mathcal{M}'$ is small and it contains the true model and models that have non-negligible values of $r_\gamma(M)$. To construct $\mathcal{M}'$, they first apply the sure independence screening (SIS) procedure of Fan and Lv (2008) and then apply LASSO and/or SCAD to those p' predictors that survived SIS, and take all those models that lie on the solution path as $\mathcal{M}'$. Note that constructing $\mathcal{M}'$ in this way will ensure the true model is captured in $\mathcal{M}'$ with high probability (Fan and Lv, 2008).

Lai *et al.* (2015) show good properties of the GFI solution both by simulation and theoretical considerations. In particular they prove a consistency theorem that we restate here.

Let M be any model, M_0 be the true model, and $\boldsymbol{H}_M$ be the projection matrix of $\boldsymbol{X}_M$; i.e., $\boldsymbol{H}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^\top \boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^\top$. Define $\Delta_M = ||\boldsymbol{\mu} - \boldsymbol{H}_M\boldsymbol{\mu}||^2$, where $\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}_{M_0}\boldsymbol{\beta}_{M_0}$.

Throughout this subsection we assume the following identifiability condition holds:

$$\lim_{n\to\infty} \min \left\{ \frac{\Delta_M}{|M_0| \log p} : \ M_0 \not\subset M, |M| \leq k|M_0| \right\} = \infty \tag{14}$$

for some fixed $k > 1$. Condition (14) is closely related to the sparse Riesz condition (Zhang and Huang, 2008).

Let $\mathcal{M}$ be the collection of models such that $\mathcal{M} = \{M : \ |M| \leq k|M_0|\}$ for some fixed k. The restriction $|M| \leq k|M_0|$ is imposed because in practice we only consider models with size comparable with the true model.

If p is large, a variable screening procedure to reduce the size is still needed. This variable screening procedure should result in a class of candidate models $\mathcal{M}'$ which satisfies

$$P(M_0 \in \mathcal{M}') \to 1 \quad \text{and} \quad \log(m'_j) = o(j \log n), \tag{15}$$

where $\mathcal{M}'_j$ contains all models in $\mathcal{M}'$ that are of size j, and m'_j is the number of models in $\mathcal{M}'_j$. The first condition in (15) guarantees the model class contains the true model, at least asymptotically. The second condition in (15) ensures that the size of the model class is not too large. The authors report small sample performance preferable to competing methods as determined by simulation study and prove asymptotic consistency of the fiducial model selection algorithm.

4 GFI for Discrete and Interval Data

Most of the material presented in Sections 2 and 3 was developed for exactly observed continuous distributions. This section discusses discrete and discretized observations.

When the observations are discrete then there is no problem with the Borel paradox and the limiting distribution in (2) can be easily computed; see Remark 2.3. In particular if we define $Q_{\boldsymbol{y}}(\boldsymbol{u}) = \{\boldsymbol{\theta} : \boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\}$ the GFD is the conditional distribution

$$V[Q_{\boldsymbol{y}}(\boldsymbol{U}^\star)] \mid \{Q_{\boldsymbol{y}}(\boldsymbol{U}^\star) \neq \emptyset\}, \tag{16}$$

where $V[A]$ selects a (possibly random) element of the closure of the set $\bar{A}$ and $\boldsymbol{U}^\star$ is an independent copy of $\boldsymbol{U}$. If $A = (a, b)$ is a finite interval, then we recommend a rule that selects one of the end points a or b at random independently of $\boldsymbol{U}^\star$ (Hannig, 2009). This selection maximizes the variance of the GFD and has been also called "half correction" (Efron, 1998; Schweder and Hjort, 2002; Hannig and Xie, 2012) and is closely related to the well-known continuity correction used in normal approximations.

4.1 Some Common Discrete Distributions

In this subsection we compute the GFDs for parameters of several popular discrete distributions.

Example 4.1. Let X be a random variable with distribution function $F(y|\theta)$. Assume there is $\mathcal{Y}$ so that $P_\theta(Y \in \mathcal{Y}) = 1$ for all θ, and for each fixed $y \in \mathcal{Y}$ the distribution function is either a non-increasing function of θ, spanning the whole interval $(0, 1)$, or a constant equal to 1. Similarly the left limit $F(y_-|\theta)$ is also either a non-increasing function of θ spanning the whole interval $(0, 1)$, or a constant equal to 0.

Define the near inverse $F^-(a|\theta) = \inf\{y : F(y|\theta) \geq a\}$. It is well known (Casella and Berger, 2002) that if $U \sim$ U(0,1), $Y = F^-(U|\theta)$ has the correct distribution and we use this association as a data generating equation.

Next, it follows that both $Q_y^+(u) = \sup\{\theta : F(y|\theta) = u\}$ and $Q_y^-(u) = \inf\{\theta : F(y_-|\theta) = u\}$ exist and satisfy $F(y|Q_y^+(u)) = u$ and $F(y_-|Q_y^-(u)) = u$. Consequently if $\boldsymbol{U}^\star$ is an independent copy of $\boldsymbol{U}$

$$P(Q_y^+(u) \leq t) = 1 - F(y|t) \quad \text{and} \quad P(Q_y^-(u) \leq t) = 1 - F(y_-|t).$$

Finally, notice that for all $u \in (0, 1)$ the function $F^-(u|\theta)$ is non-decreasing in θ and the closure of the inverse image $\bar{Q}_y(u) = \{Q_y^-(u), Q_y^+(u)\}$. Since the condition in (16) has probability 1,

there is no conditioning and the half corrected GFD has distribution function

$$R(\theta|y) = 1 - \frac{F(y|\theta) + F(y_-|\theta)}{2}.$$

If either of the distribution function is constant we interpret it as a point mass at the appropriate boundary of the parameter space.

Analogous argument shows that if the distribution function and its left limit were increasing in θ than the half corrected GFD would have distribution function

$$R(\theta|y) = \frac{F(y|\theta) + F(y_-|\theta)}{2}.$$

Using this result we provide a list of the half corrected GFDs for three well known discrete distributions. Here we understand $\text{Beta}(0, n+1)$ and $\text{Beta}(x+1, 0)$ as the degenerate distributions (Dirac measure) on 0 and 1 respectively. Similarly we understand $\Gamma(0, 1)$ as the degenerate distribution (Dirac measure) on 0.

- $X \sim \text{Binomial}(n, p)$ with n known. GFD is the 50-50 mixture of $\text{Beta}(x+1, n-x)$ and $\text{Beta}(x, n-x+1)$ distributions c.f., Hannig (2009).

- $X \sim \text{Poisson}(\lambda)$. GFD is the the 50-50 mixture of $\text{Gamma}(x+1, 1)$ and $\text{Gamma}(x, 1)$ distributions, c.f. Dempster (2008).

- $X \sim$ Negative Binomial(r, p) with r known. GFD is the 50-50 mixture of $\text{Beta}(r, x-r+1)$ and $\text{Beta}(r, x-r)$ distributions, c.f. Hannig (2014).

Example 4.2. Next we consider $Y \sim \text{Multinomial}(n, p_1, \ldots, p_k)$, where n is known and $p_i \geq 0$, $\sum_{i=1}^k p_i = 1$ are unknown.

When the categories of the multinomial have a natural ordering, Hannig (2009) suggests to write $\boldsymbol{Y} = \sum_{i=1}^n \boldsymbol{X}_i$, $q_l = \sum_{i=1}^l p_i$ and model each X_i through the data generating equation

$$\boldsymbol{X}_i = \left(I_{(0,q_1)}(U_i), I_{[q_1,q_2)}(U_i), \cdots, I_{[q_{k-1},1)}(U_i) \right)^\top, \quad i = 1, \ldots, n,$$

where $U_1, \dots U_n$ are i.i.d. U$(0,1)$ random variables. Denote the first quadrant $\mathcal{Q} = \{\boldsymbol{q} : 0 \leq q_1 \leq \dots q_{k-1} \leq 1\}$. Hannig (2009) shows that the GFD (16) for $\boldsymbol{q}$ is given by

$$V[\{\boldsymbol{q}^\star \in \mathcal{Q} : \; U^\star_{(\sum_{j=1}^{i} y_j)} \leq q_i^\star \leq U^\star_{(1+\sum_{j=1}^{i} y_j)}, \quad i = 1 \dots, k-1\}],$$

where y_i is the ith component of the observed $\boldsymbol{y}$ and $U^\star_{(j)}$ is the jth order statistics of $U_1^\star, \dots, U_n^\star$ which is an independent copy of $\boldsymbol{U}$. The GFD for $\boldsymbol{p}$ is then obtained by a simple transformation. Hannig (2009) shows good asymptotic and small sample properties of this GFD.

A drawback of the solution above is its dependency on the ordering of the categories. Lawrence *et al.* (2009) provide a solution that does not rely on a potentially arbitrary ordering of the categories. Their approach starts from analyzing each coordinate of $\boldsymbol{Y}$ individually.

As can be seen in Example 4.1, the fiducial inversion of each coordinate when ignoring the others gives a relationship $U_i \leq p_i \leq 1$ where $U_i \sim \text{Beta}(y_i, 1)$ are independent. Additionally, the fact that $\sum_{i=1}^{k} p_i = 1$ imposes a condition $\sum_{i=1}^{k} U_i \leq 1$. Consider the following random vector with its distribution taken as the conditional distribution

$$(W_0^\star, W_1^\star, \dots, W_k^\star) \sim (1 - U_1 - \dots - U_k, U_1, \dots, U_k) \mid \{U_1 + \dots + U_k \leq 1\}.$$

A straightforward calculation shows that the vector $\boldsymbol{W}$ follows Dirichlet$(1, y_1, \dots, y_k)$ distribution. Writing $Q_{\boldsymbol{y}}(\boldsymbol{w}) = \{\boldsymbol{p} : w_i \leq p_i, \; i = 1, \dots, k\}$ the GFD is $V[Q_{\boldsymbol{y}}(\boldsymbol{W}^\star)]$.

Denote by $e_i, \; i = 1, \dots k$ the coordinate unit vectors in $\mathbb{R}^k$. Notice that the set $Q_{\boldsymbol{y}}(\boldsymbol{w})$ is a simplex with vertexes $\{(w_1, \dots, w_k) + e_i w_0, \; i = 1, \dots, k\}$. The selection rule V analogous to the half correction selects each vertex with equal probability and the GFD is an equal probability $(1/k)$ mixture of Dirichlet$(Y_1 + 1, Y_2, \dots, Y_k)$, $\dots$, Dirichlet$(Y_1, Y_2, \dots, Y_k + 1)$.

4.2 Median Lethal Dose (LD50)

Consider an experiment involving k dose levels $x_1, x_2, \dots, x_k$. Each dose level x_i is administered to n_i subjects with y_i positive responses, $i = 1, 2, \dots, k$. Assume that the relationship between

dose level x_i and the probability p_i of a positive response can be represented by the logistic-linear model, given by

$$\text{logit}(p_i) = \beta_1 x_i + \beta_0 = \beta_1(x_i - \mu).$$

where $\mu = -\beta_0/\beta_1$ represents the median lethal dose (LD50) and $\text{logit}(p_i) = \log\{p_i/(1-p_i)\}$. The parameter of interest LD50 is frequently is of interest in many applied fields. Examples include a measure toxicity of a compound in a species in quantal bioassay experiments and measure of difficulty in item response models.

There are three classical methods for estimating LD50: the delta method, Fieller's method and the likelihood ratio method. If the dose-response curve is steep relative to the spread of doses, then there may be no dose groups, or at most one dose group, with observed mortalities strictly between 0% and 100%. In such cases the maximum likelihood estimator of β_1 is not calculable and the Delta method and Fieller's method fail to provide a confidence set. Furthermore, when the standard Wald test does not reject the null hypothesis $\beta_1 = 0$, Fieller's confidence sets are either the entire real line or unions of disjoint intervals. Likewise, if the null hypothesis could not be rejected by the likelihood ratio test, the likelihood ratio confidence sets are either the entire real line or unions of disjoint intervals.

E *et al.* (2009) propose a generalized fiducial solution that does not suffer these issues. They base their inference on the following data generating equation: Let $Y_{ij}, i = 1, \ldots, k, j = 1, \ldots, n_i$ denote the j^{th} subject's response to the dose level x_i. Since Y_{ij} follows a Bernoulli distribution with success probability $p_i = \text{antilogit}(\beta_0 + \beta_1 x_i)$.

$$Y_{ij} = I_{(0,\text{antilogit}(\beta_0+\beta_1 x_i))}(U_{ij}), \quad j = 1, \ldots, n_i, \ i = 1, \ldots, k.$$

Here (β_0, β_1) are unknown parameters and U_{ij} are independent standard uniform random variables.

The GFD is well-defined using (16) and E *et al.* (2009) propose to use a Gibbs sampler to implement it. They performed a thorough simulation study showing that the generalized

fiducial method compares favorably to the classical methods in terms of coverage and median length of the confidence interval for LD(50). Moreover the generalized fiducial method performed well even in the situation when the classical methods fail. They also prove that the fiducial CIs give asymptotically correct coverage, and that the effect of discretization is negligible in the limit.

4.3 Discretized Observations

In practice most datasets are rounded off in some manner, say, by a measuring instrument or by storage on a computer. Mathematically speaking, we do not know the exact realized value $\boldsymbol{Y} = \boldsymbol{y}$. Instead we only observe an occurrence of an event $\{\boldsymbol{Y} \in A_{\boldsymbol{y}}\}$, for some multivariate interval $A_{\boldsymbol{y}} = [\boldsymbol{a}, \boldsymbol{b})$ containing $\boldsymbol{y}$ and satisfying $P_{\theta_0}(\boldsymbol{Y}^\star \in A_{\boldsymbol{y}}) > 0$, where $\boldsymbol{Y}^\star = \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}_0)$ is an independent copy of $\boldsymbol{Y}$.

For example, if the exact value of the random vector $\boldsymbol{Y}$ was $\boldsymbol{y} = (\pi, e, 1.28)$ and due to instrument precision all the values were rounded to one decimal place, our observation would be the the event $A_{\boldsymbol{y}} = [3.1, 3.2) \times [2.7, 2.8) \times [1.2, 1.3)$.

Since $P_{\theta_0}(\boldsymbol{Y}^\star \in A_{\boldsymbol{y}}) > 0$ the arguments in Remark 2.3 still apply and the formula (16) remains valid with $Q_{\boldsymbol{y}}(\boldsymbol{u}) = \{\boldsymbol{\theta} : \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}) \in \bar{A}_{\boldsymbol{y}}\}$, where $\bar{A}_{\boldsymbol{y}}$ is the closure of $A_{\boldsymbol{y}}$.

Hannig (2013) proves fiducial Bernstein-von Mises theorem for discretized data. He assumes that we observed discretized i.i.d. observations with a distribution function $F(y|\boldsymbol{\theta})$. He sets $F^-(a|\boldsymbol{\theta}) = \inf\{y : F(y|\boldsymbol{\theta}) \geq a\}$ and assumes the data generating equation

$$Y_i = F^-(U_i \mid \boldsymbol{\theta}), \quad i = 1, \ldots, n,$$

where Y_i are random variables, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a p-dimensional parameter, U_i are i.i.d. $U(0, 1)$.

We restate the main theorem in Hannig (2013) in the language of this review paper:

Theorem 4.1 (Hannig, 2013)**.** *Suppose Assumption E.1 in Appendix E holds. Then the GFD defined by (16) has the same asymptotically normal distribution and satisfies Assumption 2.1*

regardless the choice of $V[\,\boldsymbol{\cdot}\,]$. *Consequently, any collection of sets* $C_n(\boldsymbol{y}_n)$ *that in the limit becomes location invariant will form asymptotically correct confidence intervals.*

4.4 Linear Mixed Models

Despite the long history of inference procedures for normal linear mixed models, a well-performing, unified inference method is lacking. ANOVA-based methods offer, what tends to be, model-specific solutions. Bayesian methods allow for solutions to very complex models, but determining an appropriate prior distribution can be confusing.

Cisewski and Hannig (2012) propose the use of GFI for discretized linear mixed models that avoids the issues mentioned above. They start with the following data generating equation:

$$\mathbf{Y} = \mathbf{X}\beta + \sum_{i=1}^{r} \sigma_i \sum_{j=1}^{l_i} \boldsymbol{V}_{i,j} U_{i,j},$$

where $\mathbf{X}$ is a known $n \times p$ fixed-effects design matrix, β is the $p \times 1$ vector of fixed effects, $\boldsymbol{V}_{i,j}$ is the $n \times 1$ design vector for level j of random effect i, l_i is the number of levels per random effect i, σ_i^2 is the variance of random effect i, and the $U_{i,j}$ are independent and identically distributed standard normal random variables.

To compute the GFD in (16) Cisewski and Hannig (2012) design a computationally efficient modification of Sequential Monte Carlo (SMC) algorithm (Doucet *et al.*, 2001; Del Moral *et al.*, 2006; Douc and Moulines, 2008). The fiducial implementation includes a custom design resampling and modification step that greatly improves the efficiency of the SMC algorithm for this model.

Cisewski and Hannig (2012) perform a thorough simulation study showing that the proposed method yields confidence interval estimation for all parameters of balanced and unbalanced normal linear mixed models. The fiducial intervals were as good as or better than the best tailor made ANOVA based solutions for the simulation scenarios covered. In addition, for the models considered by Cisewski and Hannig (2012) and for the prior selected based on rec-

ommendations in the literature, the Bayesian interval lengths were not generally competitive with the other methods used in the study.

The authors point out that even though more variation was incorporated into the data for the generalized fiducial method due to the use of discretized data, the generalized fiducial method tended to maintain stated coverage (or be conservative) while having average interval lengths comparable or shorter than other methods even though the competing methods assumed the data are observed exactly.

5 Computational Issues

This section presents some computational challenges involved when applying GFI in practice and some possible solutions to solve these challenges.

For any given model, we recall that the GFD is defined as the weak limit in (2) and under fairly general conditions, the weak limit has a density $r(\boldsymbol{\theta}|\boldsymbol{y})$ given in (3). This density can often be used directly to form estimates and asymptotic confidence intervals for the model parameters, in a similar manner as the density of the posterior distribution in the Bayesian paradigm. Standard sampling techniques such as MCMC, importance sampling or sequential Monte Carlo have been successfully implemented; e.g., Hannig *et al.* (2006a); Hannig (2009); Hannig and Lee (2009); Wandler and Hannig (2012b); Cisewski and Hannig (2012).

The exact form of generalized fiducial density could be hard to compute. For this reason, Hannig *et al.* (2014) presented a computationally tractable solution for conducting generalized fiducial inference without knowing the exact closed form of the generalized fiducial distribution.

5.1 Evaluating the Generalized Fiducial Density via Subsampling

In some situations, even the denominator of the density $r(\boldsymbol{\theta}|\boldsymbol{y})$ becomes too complicated to evaluate directly, particularly so when the l_∞ norm is used in (2). In such situations, the func-

tion $D(\cdot)$ in (4) is a sum over all possible tuples of length p; i.e., $D(A) = \sum_{\boldsymbol{i}=(i_1,\ldots,i_p)} |\det(A)_{\boldsymbol{i}}|$. If we have n observations, there are in total $\binom{n}{p}$ number of possible tuples. If the sum cannot be simplified analytically, one is obliged to compute all $\binom{n}{p}$ terms for each tuple. Such computations can become prohibitively expensive even for moderate n and p. Appropriate approximations are required in order to evaluate the density efficiently.

If the observations are i.i.d. and l_∞ norm is used, $D\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\big|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}\right)$ is a U-statistic. Given the strong dependency of the terms in $D(\cdot)$, it seems possible to use much less than $\binom{n}{p}$ terms for approximation without loss of accuracy. Blom (1976) shows that incomplete U-statistic based on random selection of K subsamples behaves very similar to the complete U-statistic when n and K are large. On the basis of this result, Hannig (2009) and its follow-up papers we suggest to replace $D(\cdot)$ by

$$\hat{D}(A;\mathcal{I}_K) = \sum_{\boldsymbol{i}\in\mathcal{I}_K} |\det(A)_{\boldsymbol{i}}|\,,$$

where $\mathcal{I}_K$ is a random selection of K different p-tuples. Numerical simulations confirm that this approximation is very promising for a wide range of applications. In practice, a common choice of K would be in the order of hundreds. One may want to choose K keeping in mind that a small K may fail to yield a good enough approximation. On the other hand, a large value of K would cause too much computations and it may be not favorable.

In most algorithms such as an MCMC sampler, the density is repeatedly evaluated for different values of $\boldsymbol{\theta}$. We recommend to keep the same choice of $\mathcal{I}_K$ for different values of $\boldsymbol{\theta}$ to gain stability of the algorithm.

The above discussion also applies to the generalized fiducial density (11) when model selection is involved.

6 Concluding Remarks and Open Problems

After many years of investigations, the authors and collaborators have demonstrated that

GFI is a useful, and promising approach for conducting statistical inference. GFI has been validated by asymptotic theory and by simulation in numerous small sample problems. In this paper we have summarized the latest theoretical and methodological developments and applications of GFI. To conclude, we list some open and important research problems about GFI.

1. As mentioned earlier, the choice of data generating equation $\boldsymbol{G}$ in (1) is not unique for many problems. Based on our practical experience gained from simulations, GFD based intervals are usually conservative and often quite short as compared to competing methods for small sample sizes. This property is not well understood as traditional asymptotic tools (including higher order asymptotics) do not explain it. Understanding this non-asymptotic phenomenon will likely help both with deeper understanding of GFI and the optimal choice of $\boldsymbol{G}$. Although our numerical experience suggests that different choices of $\boldsymbol{G}$ only lead to small differences in practical performances, it would still be important to develop an objective method for choosing $\boldsymbol{G}$.

2. As an interesting alternative, one could modify the GFD definition (2) by adding a penalty term $p(\cdot)$ on $\boldsymbol{\theta}$ to encourage sparse solutions:

$$\lim_{\epsilon \to 0} \left[\arg\min_{\boldsymbol{\theta}^\star} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\| + p(\boldsymbol{\theta}^\star) \;\middle|\; \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{U}^\star, \boldsymbol{\theta}^\star)\| \leq \epsilon \right]. \tag{17}$$

   For example, in the context of linear regression with a l_1 penalty $p(\cdot)$, just as the lasso (Tibshirani, 1996) and Dantzig selector (Candes and Tao, 2007) do, (17) will lead to sparse solutions. However, we stress that while obtaining *sparse point* estimators through a minimization problem has become a standard technique, (17) produces *sparse distributions* on the parameter space as a result of minimization. This will allow us to better evaluate the uncertainty in the model selection procedure. For some other problems, one could also try a "bending energy" term such as the one used in thin plate spline fitting.

3. One possible way to gain a deeper philosophical understanding of GFI is to find a general

set of conditions under which GFI is in some sense an optimal data dependent distribution on the parameter space (assuming such a set exists). The work of Taraldsen and Lindqvist (2013) which provides an initial result on a connection between decision theory and fiducial inference would be a good starting point.

4. It would be interesting to investigate the performance of GFI when the data generating equation is mis-specified. For example, what would happen to the empirical confidence interval coverages if $N(0,1)$ is used as the random component when the truth is in fact t with 3 degrees of freedom?

Lastly, we hope that our contributions to GFI will stimulate the growth, usage and interest of this exciting approach for statistical inference in various research and application communities.

Acknowledgements

The authors are thankful to Yifan Cui who has found numerous typos in an earlier version of this manuscript. They are also most grateful to the reviewers and the associate editor for their constructive and insightful comments and suggestions. Hannig was supported in part by the National Science Foundation under Grant No. 1016441 and 1512893. Lee was supported in part by the National Science Foundation under Grant No. 1209226, 1209232 and 1512945.

A Proof of Theorem 2.1

Recall the data generating equation (1), and assume that $\boldsymbol{U} \in \mathbb{R}^n$ is an absolutely continuous random vector with a joint density $f_{\boldsymbol{U}}(\boldsymbol{u})$, defined with respect to the Lebesgue measure on $\mathbb{R}^n$, continuous on its support $\mathcal{U}$. We need the following assumptions.

Assumption A.1. The function $\boldsymbol{G}$ has continuous partial derivatives with respect to all variables $\theta_j,\ j = 1, \ldots, p$ and $u_i,\ i = 1, \ldots n$.

Assumption A.2. For each $\boldsymbol{y}$ and $\boldsymbol{\theta}$ there is at most one $\boldsymbol{u} \in \mathcal{U}$ so that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})$. For the observed data $\boldsymbol{y}$ there is a $\boldsymbol{\theta}$ and $\boldsymbol{u} \in \mathcal{U}$ so that $\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})$. Additionally, the determinant of the $n \times n$ Jacobian matrix

$$\det\left(\frac{\boldsymbol{d}}{\boldsymbol{du}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\right) \neq 0$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{u} \in \mathcal{U}$.

Assumption A.3. The $n \times p$ Jacobian matrix $\frac{\boldsymbol{d}}{\boldsymbol{d\theta}} G(\boldsymbol{U}, \boldsymbol{\theta})$ is of rank p.

For Part (iii) of Theorem 2.1 we will also need the following assumption.

Assumption A.4. The entries of the Jacobian matrix $\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})$ have continuous partial derivatives with respect to all variables $\theta_j,\ j = 1, \ldots, p$ and $u_i,\ i = 1, \ldots n$.

The proof of Theorem 2.1 begins here. We first derive a useful formula for the likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$. Consider the implicit function

$$\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}) = 0. \tag{18}$$

If for a fixed $\boldsymbol{y}$ and $\boldsymbol{\theta}$ there is $\boldsymbol{u}$ solving (18), the implicit function theorem using Assumptions A.1 and A.2 implies that there is a neighborhood of $(\boldsymbol{y}, \boldsymbol{u})$ on which the function $\boldsymbol{u}(\boldsymbol{y})$ is uniquely defined. Moreover the function $\boldsymbol{u}(\boldsymbol{y})$ is continuously differentiable and simple calculation shows that on this neighborhood the Jacobian matrix

$$\frac{\boldsymbol{du}(\boldsymbol{y})}{\boldsymbol{dy}} = \left(\frac{\boldsymbol{d}}{\boldsymbol{du}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\right)^{-1}\Bigg|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}.$$

Consequently, since by Jacobian transformation theorem

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = f_U(\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta}))\left|\det\left(\frac{\boldsymbol{du}(\boldsymbol{y})}{\boldsymbol{dy}}\right)\right|,$$

On the other hand, if for a fixed $\boldsymbol{y}$ there is no solution $\boldsymbol{u}$ then $f(\boldsymbol{y}|\boldsymbol{\theta}) = 0$ by definition. In any case

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \left. \frac{f_U(\boldsymbol{u})}{\left|\det\left(\frac{\boldsymbol{d}}{\boldsymbol{du}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\right)\right|} \right|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})}.$$

Part (i): This is a special case of Part (ii).

Part (ii): For each $1 \leq i_1 < \cdots < i_p \leq n$ define a multi-index $\boldsymbol{i} = \{i_1, \ldots, i_p\}$ and a vector $\boldsymbol{y_i} = (y_{i_1}, \ldots, y_{i_p})$. Next define the complement multi-index $\boldsymbol{i}^{\complement} = \{i, i \notin \boldsymbol{i}\}$ and its corresponding vector $\boldsymbol{y}_{\boldsymbol{i}^{\complement}}$. Let us now consider the implicit function $\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{y}_{\boldsymbol{i}^{\complement}})$ defined by (18) with $\boldsymbol{y_i}$ held fixed at the observed values.

Fix $\boldsymbol{u} = \boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})$. If the determinant of the $p \times p$ matrix obtained by keeping only rows $\boldsymbol{i} = (i_1, \ldots, i_p)$ of the $n \times p$ Jacobian matrix

$$\det\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\right)_{\boldsymbol{i}} \neq 0, \tag{19}$$

then a direct use of implicit function theorem shows that the $n \times n$ Jacobian matrix

$$\frac{\boldsymbol{du}(\boldsymbol{\theta}, \boldsymbol{y}_{\boldsymbol{i}^{\complement}})}{\boldsymbol{d\theta}\boldsymbol{y}_{\boldsymbol{i}^{\complement}}} = \left(\frac{\boldsymbol{d}}{\boldsymbol{du}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\right)^{-1} \left. \left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta}) \ , \ \frac{\boldsymbol{dy}}{\boldsymbol{dy}_{\boldsymbol{i}^{\complement}}}\right) \right|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y},\boldsymbol{\theta})},$$

where the last matrix is obtained by concatenating the columns of the $n \times p$ and $n \times (n-p)$ Jacobian matrices on either side of the vertical line. Consequently, the joint density of the random vector $(\boldsymbol{\theta}, \boldsymbol{Y}_{\boldsymbol{i}^{\complement}})$ evaluated at the observed value $\boldsymbol{y}_{\boldsymbol{i}^{\complement}}$ is

$$h_{\boldsymbol{i}}(\boldsymbol{\theta}, \boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) \left|\det\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\right)_{\boldsymbol{i}}\right|.$$

Hannig (2013) shows that the fiducial density can be computed as proportional to the sum of the joint densities

$$r(\boldsymbol{\theta}|\boldsymbol{y}) \propto \sum_{\boldsymbol{i}=(i_1,\ldots,i_p)} h_{\boldsymbol{i}}(\boldsymbol{\theta}, \boldsymbol{y}),$$

taken as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ fixed at the observed values. There is a caveat that if for some $\boldsymbol{i}$ (19) is not satisfied, the term corresponding to that $\boldsymbol{i}$ is missing from the sum as it is of a lower order in the calculation of the fiducial density. Assumption A.3 guarantees that there is

at least one term not missing and the formula is still formally true with zeros substituted for the missing terms. The statement of Part (ii) follows.

Part (iii) Again, fix the value $\boldsymbol{u} = \boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})$. Consider the singular value decomposition

$$\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}) = A(\boldsymbol{u}, \boldsymbol{\theta})S(\boldsymbol{u}, \boldsymbol{\theta})B(\boldsymbol{u}, \boldsymbol{\theta}).$$

Here $A(\boldsymbol{u}, \boldsymbol{\theta})$ and $B(\boldsymbol{u}, \boldsymbol{\theta})$ are $n \times n$ and $p \times p$ unitary matrices respectively. $S(\boldsymbol{u}, \boldsymbol{\theta})$ is a matrix with non-negative singular values on the main diagonal and zeros everywhere else. In fact Assumption A.3 implies that the singular values are all positive.

Due to Assumption A.4 this equality can be extended to a small neighborhood of $(\boldsymbol{u}, \boldsymbol{\theta})$ so that the matrix $A(\boldsymbol{u}, \boldsymbol{\theta})$ is unitary entry-wise continuously differentiable on this neighborhood and $S(\boldsymbol{u}, \boldsymbol{\theta})$ has non-negative entries on diagonal and zeros everywhere else but the entries might no longer be in a decreasing order.

Let $\boldsymbol{\theta}(\boldsymbol{u}) = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\|_2$. Fix $\boldsymbol{y}$ at the observed value and define $\boldsymbol{x}$ through an implicit equation

$$A^\top(\boldsymbol{u}, \boldsymbol{\theta}(\boldsymbol{u}))\{\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}(\boldsymbol{u}))\} - \boldsymbol{x} = 0 \tag{20}$$

Notice that it follows from definition of l^2 projection that $(x_1, \ldots, x_p) = 0$. Furthermore if we set $\boldsymbol{x}_{\complement} = (x_{p+1}, \ldots, x_n)^\top$ we have $\|\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}(\boldsymbol{u}))\|_2 = \|\boldsymbol{x}_{\complement}\|_2$.

We now want to find the density of the random vector $(\boldsymbol{\theta}(\boldsymbol{u}), \boldsymbol{x}_{\complement})$ defined by (20) and evaluated at $\boldsymbol{x}_{\complement} = 0$. By the implicit function theorem there is a neighborhood of $(\boldsymbol{\theta}, 0)$ where $\boldsymbol{u}(\boldsymbol{\theta}, \boldsymbol{x}_{\complement})$ is one to one. The $n \times n$ Jacobian matrix evaluated at $\boldsymbol{x}_{\complement} = 0$ can be directly computed after observing that $\boldsymbol{x}_{\complement} = 0$ implies $\boldsymbol{y} - \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}(\boldsymbol{u})) = 0$:

$$\frac{\boldsymbol{du}(\boldsymbol{\theta}, \boldsymbol{x}_{\complement})}{\boldsymbol{d\theta x}_{\complement}}\Big|_{\boldsymbol{x}_{\complement}=0} = \left(\frac{\boldsymbol{d}}{\boldsymbol{du}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta})\right)^{-1} A(\boldsymbol{u}, \boldsymbol{\theta}) \left(A^\top(\boldsymbol{u}, \boldsymbol{\theta})\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}) \ , \ \frac{\boldsymbol{dx}}{\boldsymbol{dx}_{\complement}}\right)\Bigg|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})}.$$

Finally denote the first p columns of $A(\boldsymbol{u}, \boldsymbol{\theta})$ by $A_1(\boldsymbol{u}, \boldsymbol{\theta})$. Direct calculation shows that the joint density of $(\boldsymbol{\theta}(\boldsymbol{u}), \boldsymbol{x}_{\complement})$ evaluated at $\boldsymbol{x}_{\complement} = 0$ is

$$h_2(\boldsymbol{\theta}, 0) = f(\boldsymbol{y}|\boldsymbol{\theta}) \left|\det\left(A_1^\top(\boldsymbol{u}, \boldsymbol{\theta})(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}))\right)\right|_{\boldsymbol{u}=\boldsymbol{G}^{-1}(\boldsymbol{y}, \boldsymbol{\theta})}.$$

Moreover the properties of singular value decomposition imply that

$$\left|\det\left(A_1^\top(\boldsymbol{u},\boldsymbol{\theta})(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta}))\right)\right| = \sqrt{\det\left(\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\right)^\top\left(\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\boldsymbol{G}(\boldsymbol{u},\boldsymbol{\theta})\right)\right)}.$$

Finally a straightforward calculation using continuity implies that the limiting GFD in (2) is $r(\boldsymbol{\theta}|\boldsymbol{y}) \propto h_2(\boldsymbol{\theta}, 0)$ and the result follows.

Remark A.1. Similar calculation can be done also for the l_1 norm. The minimizer $\boldsymbol{\theta}(\boldsymbol{u})$ of the l_1 norm will have some of its p coordinates of the $\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{\theta}(\boldsymbol{u}))$ exactly equal to some p coordinates of $\boldsymbol{y}$. Therefore we may formulate an equation similar to (20) with A being a row permutation of an identity matrix. The final formula will be similar to the l_∞ norm with an additional term depending on KKT conditions indicating if a particular corner of the l_1 ball associated with $\boldsymbol{y_i}$ and a particular quadrant is feasible as a minimizer of the l_1 norm in (2).

B Assumptions of Theorem 2.2

Sonderegger and Hannig (2014) prove their version of the Bernstein-von Mises theorem using the l_∞ version of the Jacobian (4). In particular they have $J(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{\boldsymbol{i}} J_0(\boldsymbol{y_i}, \boldsymbol{\theta})$, where the exact form is given in part (iii) of Theorem 2.1. We will discuss other Jacobian forms at the end of this section.

We start by reviewing the standard conditions sufficient to prove asymptotic normality of the maximum likelihood estimators (Lehmann and Casella, 1998).

Assumption B.1. There are seven parts:

1. The distributions $P_{\boldsymbol{\theta}}$ are distinct.

2. The set $\{y : f(y|\boldsymbol{\theta}) > 0\}$ is independent of the choice of $\boldsymbol{\theta}$.

3. The data $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ are iid with probability density $f(\cdot|\boldsymbol{\theta})$.

4. There exists an open neighborhood about the true parameter value $\boldsymbol{\theta}_0$ such that all third partial derivatives $\left(\partial^3/\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k\right) f(\boldsymbol{y}|\boldsymbol{\theta})$ exist in the neighborhood, denoted by $B(\boldsymbol{\theta}_0, \delta)$.

5. The first and second derivatives of $L(\boldsymbol{\theta}, y) = \log f(y|\boldsymbol{\theta})$ satisfy

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}_j}L(\boldsymbol{\theta}, y)\right] = 0$$

and

$$I_{j,k}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}_j}L(\boldsymbol{\theta}, y)\cdot\frac{\partial}{\partial\boldsymbol{\theta}_k}L(\boldsymbol{\theta}, y)\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}L(\boldsymbol{\theta}, y)\right].$$

6. The information matrix $I(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \delta)$.

7. There exists functions $M_{jkl}(\boldsymbol{y})$ such that

$$\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\delta)}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_l}L(\boldsymbol{\theta}, y)\right| \leq M_{j,k,l}(y) \quad \text{and} \quad E_{\boldsymbol{\theta}_0}M_{j,k,l}(Y) < \infty.$$

Next we state conditions sufficient for the Bayesian posterior distribution to be close to that of the MLE (van der Vaart, 1998; Ghosh and Ramamoorthi, 2003). The prior $\pi(\boldsymbol{\theta})$ used is the limit of Jacobians from Assumption B.4.

Assumption B.2. Let $L_n(\boldsymbol{\theta}) = \sum L(\boldsymbol{\theta}, Y_i)$.

1. For any $\delta > 0$ there exists $\epsilon > 0$ such that

$$P_{\boldsymbol{\theta}_0}\left\{\sup_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0,\delta)}\frac{1}{n}\left(L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0)\right) \leq -\epsilon\right\} \to 1.$$

2. $\pi(\boldsymbol{\theta})$ is positive at $\boldsymbol{\theta}_0$.

Finally we state assumptions on the Jacobian function.

Assumption B.3. For any $\delta > 0$

$$\inf_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0,\delta)}\frac{\min_{\boldsymbol{i}=(i_1,\ldots,i_p)}L(\boldsymbol{\theta}, \boldsymbol{Y_i})}{|L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0)|} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0,$$

where $L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(x_i|\boldsymbol{\theta})$ and $B(\boldsymbol{\theta}_0, \delta)$ is a neighborhood of diameter δ centered at $\boldsymbol{\theta}_0$.

Assumption B.4. There is a normalization c_n so that the Jacobian function $c_n^{-1} J(\boldsymbol{Y}, \boldsymbol{\theta}) \stackrel{a.s.}{\to} \pi(\boldsymbol{\theta})$ as $n \to \infty$ uniformly on compacts in $\boldsymbol{\theta}$.

Assumption B.4 can be verified in the one-parameter case using the classical Uniform Strong Law of Large Numbers (van der Vaart, 1998; Ghosh and Ramamoorthi, 2003).

In the multi-parameter case Jacobian function $J(\boldsymbol{Y}, \boldsymbol{\theta}) = \sum_{\boldsymbol{i}} J_0(\boldsymbol{Y}_{\boldsymbol{i}}, \boldsymbol{\theta})$ is a U-statistic and the uniform convergence follows from Yeo and Johnson (2001) with $c_n = \binom{n}{p}$. In particular, Assumption B.4 is implied by Assumption B.5 below.

Assumption B.5. Let $\boldsymbol{j}$ be a multi-index with values in $\{1, 2, \ldots p\}$ and denote a vector $\boldsymbol{y}_{\boldsymbol{j}} = (y_{i_1}, \ldots, y_{i_k})$. Next define the complement multi-index $\boldsymbol{j}^{\complement} = \{i, i \notin \boldsymbol{j}\}$ and its corresponding vector $\boldsymbol{y}_{\boldsymbol{j}^{\complement}}$.

1. There exists a symmetric function $g(\cdot)$ integrable with respect to $P_{\boldsymbol{\theta}_0}$, and compact space $\bar{B}(\boldsymbol{\theta}_0, \delta)$ such that for $\boldsymbol{\theta} \in \bar{B}(\boldsymbol{\theta}_0, \delta)$ and $\boldsymbol{y} \in \mathbb{R}^p$ then $|J_0(\boldsymbol{y}; \boldsymbol{\theta})| \leq g(\boldsymbol{y})$.

2. There exists a sequence of measurable sets S_M^p such that

$$P\left(\mathbb{R}^p - \cup_{M=1}^{\infty} S_M^p\right) = 0,$$

3. For each M and for all $\boldsymbol{j}$,

$$J_{\boldsymbol{j}}\left(\boldsymbol{y}_{\boldsymbol{j}}; \boldsymbol{\theta}\right) = E_{\boldsymbol{\theta}_0}\left[J_0\left(\boldsymbol{y}_{\boldsymbol{j}}, \boldsymbol{Y}_{\boldsymbol{j}^{\complement}}; \boldsymbol{\theta}\right)\right].$$

   is equicontinuous in $\boldsymbol{\theta} \in \bar{B}(\boldsymbol{\theta}_0, \delta)$ for $\{\boldsymbol{y}_{\boldsymbol{j}}\} \in S_M^{\boldsymbol{j}}$ where $S_M^p = S_M^{\boldsymbol{j}} \times S_M^{\boldsymbol{j}^{\complement}}$.

B.1 Extension to More General Jacobians

Notice that when $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ are i.i.d., the l_∞ Jacobian satisfies $\int J_0(\boldsymbol{y}_{\boldsymbol{i}}, \boldsymbol{\theta}) f(\boldsymbol{y}_{\boldsymbol{i}}|\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$ for all $\boldsymbol{i} = (i_1 < \cdots < i_p)$.

For other version of the Jacobian we need to additionally assume that there are $J(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{\boldsymbol{i}} \tilde{J}_{\boldsymbol{i}}(\boldsymbol{y}, \boldsymbol{\theta})$, and a constant C so that $\int \tilde{J}_{\boldsymbol{i}}(\boldsymbol{y}_{\boldsymbol{i}}, \boldsymbol{\theta}) f(\boldsymbol{y}_{\boldsymbol{i}}|\boldsymbol{\theta}) d\boldsymbol{\theta} \leq C$ for all $\boldsymbol{i} = (i_1 < \cdots < i_p)$.

Finally, we remark that Assumption B.4 becomes relatively easier to verify when considering the l_2 Jacobian from part (ii) of Theorem 2.1, as one can use the Uniform Law of Large Numbers instead of uniform convergence of U-statistics.

C Proof of Theorem 2.3

Proof. Using Skorokhod's representation theorem we can assume that there is a version of data so that $t_n(\boldsymbol{Y}_n) \to \boldsymbol{T}$ almost surely. The theorem is then proved in three steps.

We need to compute

$$P(\boldsymbol{\theta}_{n,0} \in C_n(\boldsymbol{Y}_n)) = P(\boldsymbol{\xi}_0 \in \Xi_n(\boldsymbol{Y}_n)) \geq P(\boldsymbol{\xi}_0 \in C(\boldsymbol{T})) - P\left(\boldsymbol{\xi}_0 \in C(\boldsymbol{T}) \setminus \bigcap_{k=n}^{\infty} \boldsymbol{\Xi}_k(C_k(\boldsymbol{Y}_k))\right).$$

Since the set $C(\boldsymbol{t}) \setminus \bigcap_{k=m}^{\infty} \Xi_n(C_n(\boldsymbol{Y}_n))$ shrinks monotonically to $\emptyset$, $\boldsymbol{\xi}_0$ will be excluded eventually almost surely and the last probability in the equation above goes to zero. Analogously

$$P(\boldsymbol{\theta}_{n,0} \in C_n(\boldsymbol{Y}_n)) \leq P(\boldsymbol{\xi}_0 \in C(\boldsymbol{T})) + P\left(\boldsymbol{\xi}_0 \in \bigcup_{k=n}^{\infty} \boldsymbol{\Xi}_k(C_k(\boldsymbol{Y}_k)) \setminus C(\boldsymbol{T})\right).$$

By combining these two inequalities we get $P(\boldsymbol{\theta}_{n,0} \in C_n(\boldsymbol{Y}_n)) \to P(\boldsymbol{\xi}_0 \in C(\boldsymbol{T}))$.

Next, define $V_{\boldsymbol{t}}(\boldsymbol{\xi}) = \{\boldsymbol{v} : \ \boldsymbol{t} = \boldsymbol{H}(\boldsymbol{v}, \boldsymbol{\xi})\}$. Notice that $V_{\boldsymbol{t}}(C(\boldsymbol{t})) = \mathcal{V}_{\boldsymbol{t}_2}$ defined in the statement of the Theorem 2.3. Also by invertibility $V_{\boldsymbol{T}}(\boldsymbol{\xi}_0)$ has the same distribution as $\boldsymbol{V}$ in the limiting data generating equation. Recall that the limiting GFD $R_{\boldsymbol{t}}$ is the conditional distribution $Q_{\boldsymbol{t}_1}(\boldsymbol{V}_1^\star) \mid \boldsymbol{H}_2(\boldsymbol{V}_2^\star) = \boldsymbol{t}_2$, where $Q_{\boldsymbol{t}_1}(\boldsymbol{v}_1) = \boldsymbol{\xi}$ is the solution of $\boldsymbol{t}_1 = \boldsymbol{H}_1(\boldsymbol{v}_1, \boldsymbol{\xi})$. Consequently,

$$P(\boldsymbol{\xi}_0 \in C(\boldsymbol{T})) = P(V_{\boldsymbol{T}}(\boldsymbol{\xi}_0) \in \mathcal{V}_{\boldsymbol{T}_2}) = EP((\boldsymbol{V}_1, \boldsymbol{V}_2) \in \mathcal{V}_{\boldsymbol{T}_2} | H_2(\boldsymbol{V}_2) = \boldsymbol{T}_2) = E[R_{\boldsymbol{T}}(C(\boldsymbol{T}))],$$

where the second equality follows from the fact that conditionally on $\boldsymbol{T}_2 = \boldsymbol{t}_2$ the set $\mathcal{V}_{\boldsymbol{t}_2}$ is a fixed (non-random) set.

We conclude by showing that $R_{\boldsymbol{t}}(C(\boldsymbol{t})) = \alpha$ for almost all $\boldsymbol{T} = \boldsymbol{t}$. Denote $A_m = \bigcap_m^{\infty} \Xi_n(C_n(\boldsymbol{y}_n))$ and $B_m = \bigcup_m^{\infty} \Xi_n(C_n(\boldsymbol{y}_n))$. By our assumptions $R_{\boldsymbol{t}}(\partial A_m) = R_{\boldsymbol{t}}(\partial B_m) = 0$ and consequently we have $\alpha \geq \lim_{n\to\infty} R_{n,\boldsymbol{y}_n} \Xi_n^{-1}(A_m) = R_{\boldsymbol{t}}(A_m) \to R_{\boldsymbol{t}}(C(\boldsymbol{t}))$ and $\alpha \leq \lim_{n\to\infty} R_{n,\boldsymbol{y}_n} \Xi_n^{-1}(B_m) = R_{\boldsymbol{t}}(B_m) \to R_{\boldsymbol{t}}(C(\boldsymbol{t}))$ by continuity of measure. The statement of the theorem follows. □

D Proof of Theorem 3.1

We will study GFD defined for a finite collection of models $\mathcal{M}$. Recall that the data generating equation is

$$\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\theta}_M, \boldsymbol{U}), \qquad M \in \mathcal{M},\ \boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M,$$

where $\boldsymbol{y}$ is the observations, M is the model considered, $\boldsymbol{\theta}_M$ are the parameters associated with model M, and $\boldsymbol{U}$ is a random vector of with fully known distribution independent of any parameters. To derive GFD in this context we will apply definition in (2).

We start by stating an assumption closely related to identifiability.

Assumption D.1. For any two models $M_1 \neq M_2 \in \mathcal{M}$,

$$P\left(\bigcap_{i=1,2}\{\min_{\boldsymbol{\theta}_{M_i},\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}(M_i,\boldsymbol{\theta}_{M_i},\boldsymbol{U})\|\leq\epsilon\}\right) = o\left(\max_{i=1,2}P(\min_{\boldsymbol{\theta}_{M_i},\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}(M_i,\boldsymbol{\theta}_{M_i},\boldsymbol{U})\|\leq\epsilon)\right), \quad \epsilon\to 0. \quad (21)$$

A simple calculation applying the inclusion and exclusion formula to (2) gives the following result.

Lemma D.1. *Under Assumption D.1 the marginal fiducial distribution for each $M \in \mathcal{M}$ is the limit, as $\epsilon \to 0$, of the conditional probabilities*

$$r(M|\boldsymbol{y}) = \lim_{\epsilon\to 0}\frac{P(\min_{\boldsymbol{\theta}_M,\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}(M,\boldsymbol{\theta}_M,\boldsymbol{U})\|\leq\epsilon)}{x}\sum_{M'\in\mathcal{M}}P(\min_{\boldsymbol{\theta}_{M'},\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}(M',\boldsymbol{\theta}_{M'},\boldsymbol{U})\|\leq\epsilon). \quad (22)$$

We are now ready to prove Theorem 3.1. In the rest of this section we also suppose the assumptions of Theorem 2.1 hold. First notice that the invertibility implies $|M| \leq n$, where $|M|$ is the number of parameters in M. More importantly, as $\epsilon \to 0$

$$P\left(\min_{\boldsymbol{\theta}_M,\sigma^2}\|\boldsymbol{y}-\boldsymbol{G}(M,\boldsymbol{\theta}_M,\boldsymbol{U})\|_\infty\leq\epsilon\right)\sim C_M(\boldsymbol{y})\epsilon^{\min(0,n-|M|)},$$

where $C_M(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}_M} f_M(\boldsymbol{y}, \boldsymbol{\theta}_M) J_M(\boldsymbol{y}, \boldsymbol{\theta}_M)\, d\boldsymbol{\theta}_M$. Consequently the GFD in (22) assigns positive probability only to the largest model. To solve this issue we augment for each model the data generating equation $\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\theta}_M, \boldsymbol{U})$ by

$$p_k = P_k, \quad k = 1, \ldots, |M|,$$

where P_i are i.i.d. continuous random variables with $f_P(0) = q$ independent of $\boldsymbol{U}$, and q is a constant determined by the penalty. Since these extra generating equations are fully synthetic we can set the observed value to $p_i = 0$. For the augmented data generating equation we get

$$P\left(\min_{\boldsymbol{\theta}_M, \sigma^2} \|\boldsymbol{y} - \boldsymbol{G}(M, \boldsymbol{\theta}_M, \boldsymbol{U})\|_\infty \leq \epsilon, \max_{i=1,\ldots,|M|} |P_i| \leq \epsilon\right) \sim C_M(\boldsymbol{y}) q^{|M|} \epsilon^n.$$

The statement of Theorem 3.1 follows.

To conclude we remark that if the original data generating equation satisfied the identifiability assumption, so does the augmented data generating equation. To see this, notice that if $|M_1| \leq |M_2|$ then the left-hand-side of (21) is multiplied by a factor of order $\epsilon^{|M_2|}$ while the terms on the right-hand-side of (21) are multiplied only by a factor of $O(\epsilon^{|M_2|})$.

E Assumptions for Theorem 4.1

Assumption E.1. This assumption has four parts:

1. Assume that $\mathbb{R}$ is partitioned into fixed intervals

$$(-\infty, a_1], (a_1, a_2], \ldots, (a_k, \infty)$$

denoting $a_0 = -\infty,\ a_{k+1} = \infty$.

The values of Y_i are observed only up to the resolution of the grid. In other words, we do not observe the realized value y_i itself, only which of the intervals it falls into; i.e., we observe $\boldsymbol{k} = (k_1, \ldots, k_n)$ so that $y_i \in (a_{k_i}, a_{k_i+1}]$ or equivalently $\boldsymbol{y} \in (\boldsymbol{a_k}, \boldsymbol{a_{k+1}}]$ with $\boldsymbol{a_k} = (a_{k_1}, \ldots, a_{k_n})$.

2. Assume that $k \geq p$. For all $j = 0, \ldots, k$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we have $P_{\boldsymbol{\theta}}(Y \in (a_j, a_{j+1})) > 0$ and $P_{\boldsymbol{\theta}}(Y = a_j) = 0$.

3. Assume $F(y|\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$ for all $y \in \{a_1, \ldots, a_k\}$; i.e., all p first order partial derivatives are continuous.

4. For all $\boldsymbol{j} = (j_1 < \cdots < j_p) \subset \{1, \ldots, k\}$ and $u_1 < \cdots < u_p$ there is a unique solution $\boldsymbol{\theta}$ of $(F(a_{j_1}|\boldsymbol{\theta}), \ldots, F(a_{j_p}|\boldsymbol{\theta})) = (u_1, \ldots, u_p)$, and the Jacobian

$$\det\left(\frac{\boldsymbol{d}(F(a_{j_1}|\boldsymbol{\theta}), \ldots, F(a_{j_p}|\boldsymbol{\theta}))}{\boldsymbol{d\theta}}\right) \neq 0.$$

References

Barnard, G. A. (1995) Pivotal Models and the Fiducial Argument. *International Statistical Reviews*, **63**, 309–323.

Bayarri, M. J., Berger, J. O., Forte, A. and García-Donato, G. (2012) Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, **40**, 1550–1577.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025 – 2035.

Berger, J. (2011) Catalog of Objective Priors. *Tech. rep.*, Duke University.

Berger, J. O. (1992) On the development of reference priors (with discussion). *Bayesian Statistics*, **4**, 35–60.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009) The formal definition of reference priors. *The Annals of Statistics*, **37**, 905–938.

— (2012) Objective Priors for Discrete Parameter Spaces. *Journal of the American Statistical Association*, **107**, 636–648.

Berger, J. O. and Pericchi, L. R. (1996) The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

— (2001) Objective Bayesian methods for model selection: introduction and comparison. In *Model selection*, vol. 38 of *IMS Lecture Notes Monogr. Ser.*, 135–207. Beachwood, OH: Inst. Math. Statist.

Berger, J. O. and Sun, D. (2008) Objective priors for the bivariate normal model. *The Annals of Statistics*, **36**, 963–982.

Birnbaum, A. (1962) On the foundations of statistical inference. *Journal of the American Statistical Association*, **57**, 269 – 326.

Blom, G. (1976) Some properties of incomplete u-statistics. *Biometrika*, **63**, 573–580.

Candes, E. and Tao, T. (2007) The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, **35**, 2313–2351.

Casella, G. and Berger, R. L. (2002) *Statistical Inference*. Pacific Grove, CA: Wadsworth and Brooks/Cole Advanced Books and Software, 2nd edn.

Chen, J. and Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.

Chiang, A. K. L. (2001) A simple general method for constructing confidence intervals for functions of variance components. *Technometrics*, **43**, 356–367.

Cisewski, J. and Hannig, J. (2012) Generalized fiducial inference for normal linear mixed models. *The Annals of Statistics*, **40**, 2102–2127.

Dawid, A. P. and Stone, M. (1982) The functional-model basis of fiducial inference. *The Annals of Statistics*, **10**, 1054–1074.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973) Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society, Series B*, **35**, 189–233.

Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B*, **68**, 411–436.

Dempster, A. P. (1966) New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, **37**, 355–374.

— (1968) A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 205–247.

— (2008) The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, **48**, 365–377.

Douc, R. and Moulines, E. (2008) Limit theorems for weighted samples with applications to sequential monte carlo methods. *The Annals of Statistics*, **36**, 2344–2376.

Doucet, A., De Freitas, N. and Gordon, N. (2001) *Sequential Monte Carlo methods in practice*. Springer.

E, L., Hannig, J. and Iyer, H. K. (2008) Fiducial Intervals for Variance Components in an Unbalanced Two-Component Normal Mixed Linear Model. *Journal of the American Statistical Association*, **103**, 854–865.

— (2009) Fiducial Generalized Confidence interval for Median Lethal Dose (LD50). Unpublished manuscript.

Edlefsen, P. T., Liu, C. and Dempster, A. P. (2009) Estimating limits from Poisson counting data using Dempster–Shafer analysis. *The Annals of Applied Statistics*, **3**, 764–790.

Efron, B. (1998) R.A.Fisher in the 21st century. *Statistical Science*, **13**, 95–122.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, **222**, 309 – 368.

— (1925) Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700 – 725.

— (1930) Inverse Probability. *Proceedings of the Cambridge Philosophical Society*, **xxvi**, 528–535.

— (1933) The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London series A*, **139**, 343–348.

— (1935) The Fiducial Argument in Statistical Inference. *The Annals of Eugenics*, **VI**, 91–98.

Fraser, A. M., Fraser, D. A. S. and Staicu, A.-M. (2009) The second order ancillary: A differential view with continuity. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, **16**, 1208–1223.

Fraser, D. and Naderi, A. (2008) Exponential models: Approximations for probabilities. *Biometrika*, **94**, 1–9.

Fraser, D., Reid, N. and Wong, A. (2005) What a model with data says about theta. *International Journal of Statistical Science*, **3**, 163–178.

Fraser, D. A. S. (1961a) On fiducial inference. *The Annals of Mathematical Statistics*, **32**, 661–676.

— (1961b) The fiducial method and invariance. *Biometrika*, **48**, 261–280.

— (1966) Structural probability and a generalization. *Biometrika*, **53**, 1–9.

— (1968) *The Structure of Inference*. New York-London-Sydney: John Wiley & Sons Inc.

— (2004) Ancillaries and Conditional Inference. *Statistical Science*, **19**, 333–369.

— (2011) Is Bayes posterior just quick and dirty confidence? *Statistical Science*, **26**, 299–316.

Fraser, D. A. S., Reid, N., Marras, E. and Yi, G. Y. (2010) Default Priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society, Series B*, **72**.

Ghosh, J. K. and Ramamoorthi, R. V. (2003) *Bayesian Nonparametrics*. Spring-Verlag.

Glagovskiy, Y. S. (2006) *Construction of Fiducial Confidence Intervals For the Mixture of Cauchy and Normal Distributions*. Master's thesis, Department of Statistics, Colorado State University.

Hannig, J. (2009) On generalized fiducial inference. *Statistica Sinica*, **19**, 491–544.

— (2013) Generalized Fiducial Inference via Discretization. *Statistica Sinica*, **23**, 489–514.

— (2014) Discussion of "On the Birnbaum Argument for the Strong Likelihood Principle" by D. G. Mayo. *Statistical Science*, **29**, 254 –258.

Hannig, J., E, L., Abdel-Karim, A. and Iyer, H. K. (2006a) Simultaneous Fiducial Generalized Confidence Intervals for Ratios of Means of Lognormal Distributions. *Austrian Journal of Statistics*, **35**, 261–269.

Hannig, J., Iyer, H. K. and Patterson, P. (2006b) Fiducial generalized confidence intervals. *Journal of American Statistical Association*, **101**, 254 – 269.

Hannig, J., Iyer, H. K. and Wang, J. C.-M. (2007) Fiducial approach to uncertainty assessment: accounting for error due to instrument resolution. *Metrologia*, **44**, 476–483.

Hannig, J., Lai, R. C. S. and Lee, T. C. M. (2014) Computational issues of generalized fiducial inference. *Computational Statistics and Data Analysis*, **71**, 849 – 858.

Hannig, J. and Lee, T. C. M. (2009) Generalized fiducial inference for wavelet regression. *Biometrika*, **96**, 847 – 860.

Hannig, J., Wang, C. M. and Iyer, H. K. (2003) Uncertainty Calculation for the Ratio of Dependent Measurements. *Metrologia*, **4**, 177–186.

Hannig, J. and Xie, M. (2012) A note on Dempster-Shafer Recombinations of Confidence Distributions. *Electrical Journal of Statistics*, **6**, 1943–1966.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian Model Averaging: A Tutorial (with discussion). *Statistical Science*, **14**, 382–417. Corrected version available at http://www.stat.washington.edu/www/research/ online/hoetingl999.pdf.

Iyer, H. K. and Patterson, P. (2002) A recipe for constructing generalized pivotal quantities and generalized confidence intervals. *Tech. rep.*, Department of Statistics, Colorado State University.

Iyer, H. K., Wang, C. M. J. and Mathew, T. (2004) Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, **99**, 1060–1071.

Jeffreys, H. (1940) Note on the Behrens-Fisher formula. *The Annals of Eugenics*, **10**, 48–51.

Lai, R. C. S., Hannig, J. and Lee, T. C. M. (2015) Generalized fiducial inference for ultra-high dimensional regression. *Journal of American Statistical Association*. To appear.

Lawrence, E., Liu, C., Vander Wiel, S. and Zhang, J. (2009) A new method for multinomial inference using Dempster-Shafer theory. Preprint.

Lee, T. C. M. (2002) Tree–based wavelet regression for correlated data using the minimum description length principle. *Australian and New Zealand Journal of Statistics*, **44**, 23–39.

Lehmann, E. L. and Casella, G. (1998) *Theory of point estimation*. New York: Springer.

Lindley, D. V. (1958) Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society, Series B*, **20**, 102–107.

Liu, Y. and Hannig, J. (2014) Generalized Fiducial Inference for Binary Logistic Item Response Models. Preprint.

Majumder, P. A. and Hannig, J. (2015) Higher order asymptotics for Generalized Fiducial Inference. Preprint.

Martin, R. and Liu, C. (2013) Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, **108**, 301 – 313.

— (2015a) Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society, Series B*, **77**, 195–217.

— (2015b) Marginal inferential models: prior-free probabilistic inference on interest parameters. *Journal of the American Statistical Association*. To appear.

Martin, R. and Walker, S. G. (2014) Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.*, **8**, 2188–2206. URL `http://dx.doi.org/10.1214/14-EJS949`.

Martin, R., Zhang, J. and Liu, C. (2010) Dempster-Shafer theory and statistical inference with weak beliefs. *Statistical Science*, **25**, 72–87.

McNally, R. J., Iyer, H. K. and Mathew, T. (2003) Tests for individual and population bioequivalence based on generalized p-values. *Statistics in Medicine*, **22**, 31–53.

Patterson, P., Hannig, J. and Iyer, H. K. (2004) Fiducial Generalized Confidence Intervals for Proportion of Conformance. *Tech. rep.*, Colorado State University.

Salome, D. (1998) *Staristical Inference via Fiducial Methods*. Ph.D. thesis, University of Groningen.

Schweder, T. and Hjort, N. L. (2002) Confidence and likelihood. *Scandinavian Journal of Statistics*, **29**, 309–332.

Singh, K., Xie, M. and Strawderman, W. E. (2005) Combining information from independent sources through confidence distributions. *The Annals of Statistics*, **33**, 159–183.

Sonderegger, D. and Hannig, J. (2014) Fiducial theory for free-knot splines. In *Contemporary Developments in Statistical Theory, a Festschrift in honor of Professor Hira L. Koul*, 155 – 189. Springer.

Stevens, W. L. (1950) Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, **37**, 117–129.

Taraldsen, G. and Lindqvist, B. H. (2013) Fiducial theory and optimal inference. *The Annals of Statistics*, **41**, 323–341.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Tsui, K.-W. and Weerahandi, S. (1989) Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, **84**, 602–607.

— (1991) Corrections: "Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters" [J. Amer. Statist. Assoc. **84** (1989), no. 406, 602–607; MR1010352 (90g:62047)]. *Journal of the American Statistical Association*, **86**, 256.

Tukey, J. W. (1957) Some examples with fiducial relevance. *The Annals of Mathematical Statistics*, 687–695.

van der Vaart, A. W. (1998) *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

Veronese, P. and Melilli, E. (2014) Fiducial and Confidence Distributions for Real Exponential Families. *Scandinavian Journal of Statistics*, in press.

Wandler, D. V. and Hannig, J. (2011) Fiducial Inference on the Maximum Mean of a Multivariate Normal Distribution. *Journal of Multivariate Analysis*, **102**, 87 – 104.

— (2012a) A fiducial approach to multiple comparisons. *Journal of Statistical Planning and Inference*, **142**, 878 – 895.

— (2012b) Generalized Fiducial Confidence Intervals for Extremes. *Extremes*, **15**, 67–87.

Wang, J. C.-M., Hannig, J. and Iyer, H. K. (2012a) Fiducial Prediction Intervals. *Journal of Statistical Planning and Inference*, **142**, 1980–1990.

— (2012b) Pivotal methods in the propagation of distributions. *Metrologia*, **49**, 382–389.

Wang, J. C.-M. and Iyer, H. K. (2005) Propagation of uncertainties in measurements using generalized inference. *Metrologia*, **42**, 145–153.

— (2006a) A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. *Measurement*, **39**, 856–863.

— (2006b) Uncertainty analysis for vector measurands using fiducial inference. *Metrologia*, **43**, 486–494.

Wang, Y. H. (2000) Fiducial intervals: what are they? *The American Statistician*, **54**, 105–111.

Weerahandi, S. (1993) Generalized confidence intervals. *Journal of the American Statistical Association*, **88**, 899–905.

— (1994) Correction: "Generalized confidence intervals" [J. Amer. Statist. Assoc. **88** (1993), no. 423, 899–905; MR1242940 (94e:62031)]. *Journal of the American Statistical Association*, **89**, 726.

— (1995) *Exact statistical methods for data analysis.* Springer Series in Statistics. New York: Springer-Verlag.

Welch, B. L. and Peers, H. W. (1963) On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society: Series B*, **25**, 318–329.

Wilkinson, G. N. (1977) On resolving the controversy in statistical inference. *Journal of the Royal Statistical Society, Series B*, **39**, 119–171.

Xie, M., Liu, R. Y., Damaraju, C. V. and Olson, W. H. (2013) Incorporating external information in analyses of clinical trials with binary outcomes. *The Annals of Applied Statistics*, **7**, 342–368.

Xie, M. and Singh, K. (2013) Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, **81**, 3 – 39.

Xie, M., Singh, K. and Strawderman, W. E. (2011) Confidence distributions and a unified framework for meta-analysis. *Journal of the American Statistical Association*, **106**, 320–333.

Xu, X. and Li, G. (2006) Fiducial inference in the pivotal family of distributions. *Science in China: Series A Mathematics*, **49**, 410–432.

Yeo, I. K. and Johnson, R. A. (2001) A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. *Statistics and Probability Letters*, **51**, 63–69.

Zhang, C. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567–1594.

Zhang, J. and Liu, C. (2011) Dempster-Shafer inference with weak beliefs. *Statistica Sinica*, **21**, 475–494.