Sub-model selection and combination for statistical inference

D.A.S. Fraser

*Department of Mathematics and Statistics, York University, Department of Statistics, University of Toronto

Abstract
Recent statistical inference has largely developed from two basic paradigms: the conditional methods of exponential models, and the marginal methods of transformation models. This paper recommends an open sequential extraction of marginal and conditional components, the interpretation of each on the basis of repeated sampling of the component, and the recombination of these for inference purposes. Examples are presented from life testing, mortality rates, and proportional hazard models.

Keywords: Conditional inference; Extraction; Life testing; Marginal inference; Mortality rates; Proportional hazard; Selection.

1. INTRODUCTION

Conditional methods for exponential models have developed from the similar test procedures of Neyman and Pearson (1933). Marginal methods for transformation models come from the proposals of Fisher (1934). Recent extensions of these are discussed in Section 2 together with a distinction between direct and pivotal inference.

Section 3 proposes a general sequential extraction of submodels, the assessment of each based on repeated sampling of the submodel, and the independence recombination of the selected components to obtain an overall inference model. Typically the components would be selected on the basis of the quality of the inference for an interest parameter, such as accuracy and freedom from nuisance parameters. Examples are discussed in Sections 4, 5, 6.

2. CONDITION AND MARGINAL INference

Conditional methods of inference have developed largely from the similar tests of Neyman and Pearson (1933). For a model with parameter \( \hat{\theta} = (\psi, \lambda) \) a similar size \( \alpha \)
test of the hypothesis $\psi = \psi_0$ satisfies $P(C; \psi_0, \lambda) = \alpha$ for all $\lambda$ where $C$ is the critical region. If there is a complete minimal sufficient statistic $s(y)$ for $\lambda$ with $\psi$ fixed at $\psi_0$, then the test satisfies $P(C|s; \psi_0) = \alpha$ for all $s$ and thus can be viewed as a conditional test given the minimal sufficient statistic for the nuisance parameter. More generally this pattern suggests that inference concerning $\psi = \psi_0$ in the presence of a nuisance parameter $\lambda$ be based on the conditional model given a minimal sufficient statistic for $(\psi_0, \lambda)$, or even given just the maximum likelihood estimate of $(\psi_0, \lambda)$.

For example the two sample, $(y_1, \ldots, y_n), (y_{n+1}, \ldots, y_{n+m})$, nonparametric problem with distribution functions $F(y), F(y - \psi)$, hypothesis $\psi = 0$, and general $F$ has minimal sufficient statistic $(y(n), \ldots, y(n+m))$ under the hypothesis. The conditional test given the order statistic is then a permutation test.

The exponential model also provides a nice example when the parameters are canonical:

$$\exp\{\psi'y_1 + \lambda'y_2 - \kappa(\psi, \lambda) + h(y)\}.$$ 

The conditional distribution of $y_1$ given $y_2$ is $\lambda$ free; thus $y_2$ is minimal sufficient for $(\psi_0, \lambda)$. This suggests that if the general factorization

$$f(y; \theta) = f(y_1|y_2; \psi)f(y_2; \psi, \lambda)$$

obtains, inference on $\psi$ be based on the first factor. The factorization that goes with the nonparametric example above is actually more general, and has a sufficient statistic $s(y) = s(y; \psi_0)$ which varies with the parameter value being tested. Some approximate methods based on the above factorization are discussed in Fraser and Reid (1988, 1989).

Certain marginal approaches to inference have developed from the location-scale analysis of Fisher (1934). For sampling with the location-scale model $\sigma^{-1} f(\sigma^{-1}(y - \mu))$ a preliminary reduction is obtained by conditioning on a configuration statistic

$$d = \hat{\sigma}^{-1}(y_1 - \hat{\mu}, \ldots, y_n - \hat{\mu})$$

giving the distribution

$$f(\hat{\mu}, \hat{\sigma} \mid d; \mu, \sigma) = \prod f(\sigma^{-1}(\hat{\mu} - \mu + \hat{\sigma} d_i)) \sigma^{-n} \hat{\sigma}^{n-2}.$$ 

Inference for $\mu$ would then be based on the marginal distribution of $t = (\hat{\mu} - \mu)/\hat{\sigma}$, and for $\sigma$ on the marginal distribution of $\hat{\sigma}/\sigma$. In each case the distribution is free of the nuisance parameter.

The preceding pattern extends to general transformation models and indicates a general pattern

$$f(y; \theta) = f(y_1; \psi)f(y_2|y_1; \psi, \lambda)$$

supporting the use of the marginal distribution of a variable $y_1$ for inference concerning $\psi$. This can include the case where $y_1$ itself depends on $\psi$; for example, above, we would use $(\hat{\mu} - \mu)/\hat{\sigma}$ to assess the interest parameter $\psi = \mu$. 
A distinction that can be useful in obtaining inference for a parameter \( \psi \) with or without a nuisance parameter centers on whether one can directly measure a parameter or only departure of data from a parameter; the first we call direct inference; and the second pivotal. A simple example is available for sampling from the normal \((\mu, \sigma)\): the estimate \( \hat{\sigma} = s_y \) directly measures \( \sigma \), and its chi-square distribution provides a full inference model; the pivot \( t = (\hat{\mu} - \mu) / \hat{\sigma} \) measures departure of an estimate \( \hat{\mu} \) from \( \mu \) but corresponds to a different partition of the basic sample space for each \( \mu \) value. It seems one would prefer direct inference when available but may need to make do with pivotal inference otherwise.

A second example where both are available involves data \((y_1, y_2)\) from the standard bivariate normal with mean \( \eta = (\rho \cos \theta, \rho \sin \theta) \) on a circle of known radius \( \rho \). Direct inference is available using the maximum likelihood \( \hat{\theta} \) given by \((y_1, y_2) = (r \cos \hat{\theta}, r \sin \hat{\theta}) \) and assessing it from its marginal distribution (the projected normal), or its conditional distribution (von Mises) given the ancillary \( r = (y_1^2 + y_2^2)^{1/2} \). Alternatively, pivotal inference is available from

\[
d = -\sin \theta y_1 + \cos \theta y_2
\]

assessed against the standard normal distribution. Theoretical preference would perhaps go to the direct analysis conditional on the radial distance \( r \).

More generally, direct and pivotal patterns provide two approaches for approximate conditional inference.

3. SUBMODEL EXTRACTION AND RECOMBINATION

We propose a widely general examination of marginal and conditional components and their recombination to produce a terminal model for inference.

As a basic unit we consider an inference unit \((\mathcal{M}, \mathcal{D})\) consisting of a statistical model \(\mathcal{M}\) and a data array \(\mathcal{D}\) recording observed values of the variables in the model. For example, we could consider the gamma model with mean \(\mu\) and shape parameter \(\beta\):

\[
\mathcal{M} : (y_1, \ldots, y_n) \sim \text{gamma}(\mu, \beta)
\]
\[
\mathcal{D} : (y_1^0, \ldots, y_n^0).
\]

An inference unit has been called an instance of evidential meaning by Birnbaum (1962) and an inference base by Fraser (1979).

We will speak of splitting an inference unit when we have a sample space partition \(\mathcal{P}\) that produces a marginal model \(\mathcal{M}_\ast\) for the variable generating the partition and a residual conditional model \(\mathcal{M}_\ast^\ast\). For example, with the variable \(y_1\) relative to \((y_1, y_2)\) we would obtain

\[
\mathcal{M}_\ast = \{f(y_1; \theta)\}, \quad \mathcal{M}_\ast^\ast = \{f(y_2 | y_1; \theta)\}
\]

and thus split the initial inference unit \((\mathcal{M}, \mathcal{D})\) into components,

\[
(\mathcal{M}_\ast, \mathcal{D}_\ast), \quad (\mathcal{M}_\ast^\ast, \mathcal{D}_\ast^\ast),
\]
based on the observed value $y^0$ of the partitioning variable. As a more general example, we might have $(y_1, y_2) \sim \text{unif}(\psi \pm \delta/2)$ and, using the partitioning variable $d = (y_1 - y_2)$, obtain

$$M_*: d \sim \text{triangular}(-\delta, +\delta)$$

$$M_*: \bar{y}|d^0 \sim \text{uniform}(\psi \pm (\delta - d^0)/2)$$

with corresponding inference units

$$(M_*, d^0), \quad (M^*, \bar{y}^0).$$

Given components obtained by splitting an inference unit, we will speak of segregating when a component is

a) selected as a terminal component for statistical inference
b) omitted as not being useful for the inference process
c) retained for further splitting and segregation.

We thus envisage a sequential process leading eventually to a collection of selected inference units as a final stage in the splitting and segregation process. Thus for example with the sample of 2 from the uniform $(\psi \pm \delta/2)$ and interest parameter $\psi$, we might select for final inference the component

$$(M^*, \bar{y}^0)$$

in this case, the nuisance parameter is still involved. We might however prefer an alternative splitting and segregating process that leads to a Student-type statistic, $t = (\bar{y} - \psi)/|d|$.

As a very general approach to statistical inference, we consider arbitrary splitting, segregating, and selection to produce a terminal collection of inference units. What criteria for this process are appropriate? How will the selected components be assessed and combined?

a) The splitting, segregating, and selection will be focussed on the quality of the inference for the interest parameter in terms of precision and accuracy, and freedom from the nuisance parameters.

b) Each inference unit model will be assessed in terms of repeated sampling of the component model situation, that is, the component model will be taken as providing the appropriate probabilities directly.

c) The component models will be combined on the basis of independence. This is the appropriate combination when the component models are taken as describing repeated sampling of the component situations; even when are not independent!

Clearly feasibility can be a criterion. Thus an analysis that is computationally accessible would be preferred to one that may be better but not computationally possible with given resources.

Also we would lose information concerning an interest parameter that is in an omitted statistical unit. We would accept the omission if the interest parameter was badly confounded with the nuisance parameter or if the information concerning the interest parameter was computationally inaccessible.
The selection of components would be based on the quality of the inference in terms of precision and unbiasedness, and not closeness to some prechosen value of interest to an investigator. Thus the example of Cox and Hinkley (1974, p. 50) would violate the selection process. The example involves sequential $N(\mu, 1)$ sampling until $|\bar{y}_n| > kn^{-1/2}$, the likelihood ratio for assessing $\mu = 0$ then approximates $\exp\{-k^2/2\}$ and thus the selection process is not based on good inference for $\mu$ but on artificial negative ‘inference’ concerning the value $\mu = 0$.

We are thus proposing a very general approach to marginal and conditional inference that can include both direct and pivotal analyses. Some examples are given in the following sections. Some of the issues considered here are relevant for meta-analysis.

4. LIFE TESTING

Life testing with exponential life and exponential censoring is commonly formalized as $n$ units under testing. For each unit we have $x_0$ distributed as $\theta_1 \exp(-\theta_1 x_0)$ and $y$ distributed as $\theta_2 \exp(-\theta_2 y)$ with observable variables

$$x = \min(x_0, y), \quad \delta = 1, 0 \sim x = x_0, y;$$

$\delta$ is $\text{Bern}(\theta_1/(\theta_1 + \theta_2))$ and $x$ is exponential with rate $(\theta_1 + \theta_2)$. For a sample of $n$ we then obtain the minimal sufficient statistic $n_0 = \Sigma \delta_i$ and $t = \Sigma x_i$ distributed independently as

$$n_0 \sim \text{Binomial}(n, \theta_1/(\theta_1 + \theta_2))$$

$$(\theta_1 + \theta_2)t \sim \text{gamma}(n).$$

The natural density factorization separates the parameter components as $\theta_1/(\theta_1 + \theta_2)$, $\theta_1 + \theta_2$ but does not isolate the parameter $\theta_1$ which is usually of interest. The likelihood function from the data,

$$L = \theta_1^{n_0} e^{-\theta_1 t} \cdot \theta_2^{n-n_0} e^{-\theta_2 t},$$

does however factor to separate the parameter $\theta_1$. The first factor is called the partial likelihood for $\theta_1$ and is often analyzed by first order asymptotics, although in the present formulation it is not a likelihood — a probability for an observed variable.

The typical context for such life testing is concerned with the basic failure rate $\theta_1$. The exponential censoring may often be quite idealistic and a general censoring distribution $f(y; \theta_2)$ more appropriate. Indeed the censoring may not be fully random and we could have immigration as well as censoring. If the failure rate $\theta_1$ is of prime interest, then we argue that the basic experimental unit is exposure and a particular investigation has some total exposure. From this viewpoint we then take the appropriate sub-model for $\theta_1$ inference to be the Poisson $(\theta_1 t)$ with model

$$f = (\theta_1 t)^x e^{-\theta_1 t}/x!$$
for the number of failures $x = n_0$. This gives a conditional analysis given a total exposure $t$.

The preceding conditional model has actual likelihood for $\theta_1$ equal to the partial likelihood mentioned earlier. Also the conditional model is an exponential family with canonical parameter $\log(\theta_1 t)$. Thus the saddlepoint approximation can be used to give accurate significance probabilities or a significance function accurate to third order.

5. MORTALITY RATES

Consider the data collected as part of an insurance program on a particular population segment. If the data were very extensive it would be possible to develop a new mortality table. With more moderate data we could reasonably expect to obtain adjustments to the currently used mortality table.

Following the pattern in Section 4 we argue that the basic experimental unit is exposure to mortality. For a given subgroup let $y_i$ be the number of deaths and $t_i$ be the corresponding exposure; also let $\mu_i$ be the mortality rate indicated by the current table. For an updated model we follow the generalized linear model pattern and assume the logarithm of the new $\nu_i$ has the form

$$\log \nu_i = \log \mu_i + \alpha + X_i \beta$$

where $\alpha$ gives a percentage overall adjustment and $\beta$ records differential coefficients for concomitant variables recorded in the row vector $X_i = (x_{i1}, \ldots, x_{ip})$ for the $i$th subgroup. For this group we condition on exposure and obtain the Poisson model

$$f = \exp\{y_i \theta_i - e^{\theta_i}\}/y_i!$$

with canonical parameter

$$\theta_i = \log(\mu_i t_i) + \alpha + X_i \beta.$$  

For the full data we then have a generalized linear model with parameters $\alpha$ and $\beta_1, \ldots, \beta_p$ recording an overall adjustment and differential adjustments with respect to the concomitant variables. The inferences can then be first order following McCullagh and Nelder (1989) or third order using saddlepoint methods as in Fraser, Reid, Wong (1991); see also Pierce and Peters (1991).

6. PROPORTIONAL HAZARDS MODELLING

The typical context for proportional hazards modelling involves objects exposed in time to the occurrence of an event $E$. The occurrence rate can depend on many factors varying in time and also on individual differences between the objects. For the $i$th object in time interval $(t, t + dt)$, the occurrence probability for $E$ has the form

$$\lambda(t) \exp \{g(X_i, \beta)\}$$
where \( \lambda(t) \) would be referred to as the base line hazard, and where \( g(X_i, \beta) \) records the logarithmic departure from the base line rate, the row vector \( X_i \) records the values of \( p \) variables describing key factors of the object, and \( \beta \) records the dependence on these factors; typically \( g(X_i, \beta) \) is taken to be linear \( X_i \beta \) as a first approximation.

The proportional hazards model was proposed by Cox (1972) and is described by Kalbfleisch and Prentice (1980) in their text on failure time analysis. Cox recommended the use of a partial likelihood

\[
L(\beta) = \prod_i \frac{\exp\{X_i\beta\}}{\sum_{j \in R(i)} \exp\{X_j\beta\}}
\]

(6.1)

where \( R(i) \) is a set of indexes for objects under exposure at the time the \( ith \) object failed and \( i \) gives summation over the objects that failed.

With an arbitrary base line hazard rate, our arguments from earlier sections would focus on the component multivariate Bernoulli model applicable at a failure time: specifically when a failure occurs it could be any of the objects then under exposure and the probabilities for the \( j \)th would be proportional to \( \exp\{X_j\beta\} \). Thus the \( ith \) factor in the likelihood \( L(\beta) \) is a Bernoulli probability for the \( ith \) object among the \( R(i) \) then exposed. We view these Bernoulli sub-models as the appropriate component models for \( \beta \) inference.

By examining and analyzing the failure events conditionally, given the failure times, we obtain a compound Bernoulli model with actual likelihood (6.1); the parameter \( \beta \) is then a canonical parameter in an exponential model and the saddlepoint methods of (Barndorff-Nielsen and Cox, 1979), Lugannani and Rice (1980) and Fraser, Reid, and Wong (1991) are applicable.

As an illustration we examine the data from Kalbfleisch and Prentice (1980, p. 82). The data records the time from insult with a carcinogen to mortality from vaginal cancer in rats under test. The rats had been divided into two groups, a control group with \( x = 0 \) and a pretreatment regime group with \( x = 1 \). The baseline hazard \( \lambda(t) \) was then adjusted to \( \lambda(t) \exp\{\beta x\} \) where \( \beta \) gives the logarithmic change due to pretreatment.

Our conditional submodel analysis is to ignore all model components except at the mortality times. Thus with the first failure at time 142 there were 19 controls, 21 treated, and with a treated rat dying, the multivariable Bernoulli probability for the particular death is then \( e^\beta/(19 + 21e^\beta) \). The second failure was at time 143 with 19 controls, 20 treated, and with a treated rat dying; the multivariable Bernoulli probability is then \( e^\beta/(19 + 20e^\beta) \). The overall data probability is obtained by compounding these components. Some rats were censored and do not contribute of course to the compound probability.

For an inference summary we focus on the significance function

\[
p(\beta) = P(\hat{\beta} \leq \beta^0; \beta)
\]

recording the left tail probability at the data point as a function of the pretreatment parameter \( \beta \). The log-likelihood function is plotted in Figure 1 and various approximations to the significance function for larger values of \( \beta \) are given in Figure 2.
The observed maximum likelihood value is −0.57 suggesting a mortality reduction due to pretreatment. The 95 percent confidence intervals and significance for β = 0, by the first order methods of maximum likelihood and signed likelihood ratio and the third order method of saddlepoint, are as follows:

\[
\begin{align*}
\text{mle} & \quad (-1.27, 0.10) & 0.045 \\
\text{sllr} & \quad (-1.254, 0.121) & 0.052 \\
3\text{rd} & \quad (-1.253, 0.117) & 0.051
\end{align*}
\]

With most exponential models the third order method is highly accurate, the likelihood ratio is often close, and the maximum likelihood and score are often inaccurate with opposite errors from the exact. In this example the sample size is large thus making first order methods quite accurate. Kalbfleisch and Prentice use an approximate likelihood and obtain the maximum likelihood value −0.60, but the difference from the present value corresponds essentially to the approximation. We recommend the conditional analysis with the available third order accuracy.

7. CONCLUDING REMARKS

We have proposed a very general search for and selection of submodels in an investigation. The object is to better measuring the parameters of interest and to eliminate submodels that contain confounded information. The control criterion is the quality of the inference for the interest parameter. Probabilities are in terms of repeated sampling of the components and the independence recombination of these.

Historically there is a well pursued pattern of choosing marginal variables that estimate or produce inferences for parameters of interest; distribution for these are possible by theory, by Monte Carlo, and by resampling methods. Our proposals extend this to the conditional context and to the combination of marginalization and conditioning. We are not interested in component investigations that were not performed, only in the actual component investigations and the repeated sampling probabilities from these.

There may be some risks with this extreme approach, but they are not apparent at this point. Clearly one cannot choose the nonparametric test that gives the answer one wants. So clearly here we do not choose components to obtain an answer one wants. Rather we choose to seek quality of inference.

REFERENCES


Figure 1. The log-likelihood for $\beta$ with the Pike data.
Figure 2. The significance function $P(\beta)$ for $\beta > 0$ as approximated by the score --- maximum likelihood ..., likelihood ratio --- (first order methods), and saddlepoint --- (third order).