Introduction.

- This paper proposes a new method to test fit of an unobserved (latent) variable to a given distribution based on an observed variable and the connection between the two.
- This is largely an unexplored area of Statistics. Here Bayesian principles are borrowed to generate the test procedure.
- By choosing an appropriate prior, this procedure yields the Cramér-von Mises or the Anderson Darling as test statistics.
- The test has good power against many alternative densities. Power study results are available upon request.
- The procedure was motivated by a paper by Krumbien (1935) and Krumbein's data will be used to give the first (simple) illustration.
- For a modern example, a test will be applied to test the distribution of the frailty term in a survival analysis model. Shih and Louis (1995) showed that different frailty distribution induce quite different dependence and therefore testing frailty distribution is necessary.

Theory of the Test

- Observed variables:$x = [x_1, \cdots, x_n]'$
- Unobserved variables: $r = [r_1, \cdots, r_n]'$
- H_o : r follows density $g(r|\theta)$, with cdf $G(r|\theta)$; prior for parameter θ is $\pi(\theta)$.
- H_A: Alternative density for r: $g(r|\theta)\Lambda(r|\theta)$, where $\Lambda(r|\theta)$ is a prior distribution to model departure from the null distribution.
- The Prior density $\Lambda(r|\theta)$ provides a set of possible alternative distributions that form an alternative band around the null density
- Conditional density of x given r: $h(x|r)$.
- The null density of x is

$$f_o(x) = \int_r \int_\theta h(x|r) g(r|\theta) \pi(\theta) d\theta dr.$$

The alternative density $f_A(x)$

$$
\begin{aligned}
f_A(x) &= \int_r \int_\theta \int_\Lambda h(x|r) g(r|\theta) \Lambda(r|\theta) \pi(\theta) d\Lambda d\theta dr \\
&= \int_r \int_\theta f(r, x, \theta) \int_\Lambda \Lambda(r|\theta) d\Lambda d\theta dr \\
&= f_o(x) \int_r \int_\theta f(r, \theta|x) \int_\Lambda \Lambda(r|\theta) d\Lambda d\theta dr \\
&= f_o(x) E\left[\int_\Lambda \Lambda(r|\theta) d\Lambda | x\right],
\end{aligned}
$$

- The Neyman-Pearson Lemma shows the likelihood ratio (LR) will maximize power subject to level α.
- The LR test statistic is $f_A(x)/f_o(x) = E\left[\int_\Lambda \Lambda(r|\theta) d\Lambda | x\right]$
- Rejection Region: $C = \{x | E[\int_\Lambda \Lambda(r|\theta) d\Lambda | x] > k\}$; when under H_o, Prob$[x \in C] = \alpha$.

Define prior density $\Lambda(r, \theta)$

- Sun and Lockhart (2018) suggest one possibility is

$$\Lambda(r|\theta) = \exp\left\{\epsilon \sum_{i=1}^{n} Z(u_i)\right\} / \left[\int_0^1 \exp\{\epsilon Z(t)\}dt\right]^n,$$

  where $u_i = G(r_i; \theta)$, the Probability Integral Transformation of r_i under the H_0

- $Z(t)$ is a Gaussian Process with $E\{Z(t)\} = 0$, $0 < t < 1$ and covariance function $\rho(s, t)$, and ϵ determines how far H_A departs from H_o.
- Classically, the Gaussian Process may be written $Z(t) = \sum_j^{\infty} w_j \sqrt{\lambda_j} g_j(t)$ where w_j are iid $N(0, 1)$ and λ_j, $g_j(t)$ are eigenvalues and eigenfunctions of

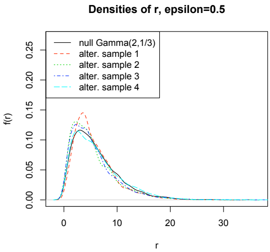
$$\int_0^1 \rho(s, t) g_j(t) dt = \lambda_j g_j(s).$$

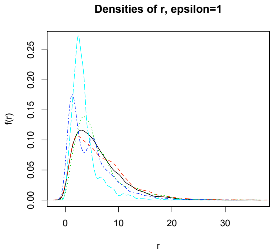
Test Statistic with the Gaussian Prior

Then the test statistic becomes

$$S = E\left(\exp\left\{\sum_{j=1}^{\infty}\frac{\epsilon^2\lambda_j[\sum_{i=1}^{n} g_j(u_i)]^2}{2(n\epsilon^2\lambda_j+1)}\right\}\middle|x\right). \quad (1)$$

when $\rho(s,t) = min(s,t) - st$, $\lambda_j = 1/(\pi^2 j^2)$ and $g_j(s) = \sqrt{2}\cos(\pi j s)$, S is the average Cramér-von Mises W^2.
Figure 1 and 2: comparisons of Null and Alter. densities.





Test Procedures

For statistic S, the posterior joint density of r and θ must be found by Gibbs sampling. The steps are as follows:

1. Assign prior distributions for θ in $g(x|\theta)$.
2. Sample from $f(r, \theta|x)$ by the Gibbs sampling algorithm: sample iteratively from the full conditionals:
   (a) $f(r_i|r^*, \theta, x)$, $i = 1, \cdots, n$ where r^* denotes the set of r omitting r_i, this gives n new values of r for a given θ
   (b) $f(\theta|r, x)$.
3. Compute S in (1) using the $[r, \theta]$ generated from the previous step (after a burn-in period), this is the test statistic S^* based on the given x values.
4. To obtain the distribution of S, bootstrap a set of $\mathbf{r}$ from the null distribution, using an estimate of θ: this may be the posterior modes in step 2 or the EM algorithm estimate or some other efficient estimates

Test Procedures

Steps continue...

5. Generate a set of $\mathbf{x}$ from $h(x|\mathbf{r})$.
6. Calculate the statistic S, call it S_d.
7. Repeat Steps $4-6$ M times, say 100, to get M values of S_d; these give an estimated distribution of S under H_o.
8. Hence calculate the p-value of S^*.

Example 1: Krumbein (1953) studied a dataset of spherical rocks. The data observed, x, are the radii of circular cross-sections of the rocks from random slicing. The variable of interest is the radii (r) of the rocks, which embedded in sediments and therefore are unobserved. Researchers wanted to test if the radii of the rocks follow a Gamma distribution, so the null hypothesis is $H_0 : \mathbf{r} \sim$ Gamma (α, β). The above steps become

1. Choose a conjugate prior density for α and β:

$$\pi(\alpha, \beta | v, q, s, t) \propto \frac{v^{\alpha-1} e^{-q/\beta}}{\Gamma(\alpha)^t \beta^{\alpha s}}.$$

Krumbein's Example.

2. A Gibbs sampler scheme is used which alternates among $\mathbf{r}, \alpha$, and β , the full conditionals for $p(r_i, \alpha, \beta | x_i, v, q, s, t)$:

$$p(r_i|x_i, \alpha, \beta) \propto \frac{r_i^{\alpha-2} e^{-r_i/\beta}}{\sqrt{r_i^2 - x_i^2}},$$

$$\pi(\alpha|\mathbf{r}) \propto \frac{(p \prod_{i=1}^n r_i / \beta^{(n+s)})^\alpha}{\Gamma(\alpha)^{n+t}},$$

$$\pi(\beta|\mathbf{r}) \propto e^{-(q+\sum_{i=1}^n r_i)/\beta} \beta^{-\alpha(n+s)}.$$

3. Calculate $S = 0.199$ in equation (1).
4. The EM algorithm is used to estimate α and β,

$$L = \prod_{i=1}^n x_i / (\Gamma(\alpha)\beta^\alpha) \int_{x_i}^{\infty} \frac{r_i^{\alpha-2} \exp(-r/\beta)}{\sqrt{r_i^2 - x_i^2}} dr_i;$$

5. The p value is 0.043. Gamma is rejected at the 5% level.

Testing a Frailty Term.

McGilchrist and Aisbett (1991) give time to first and second recurrence of kidney infection in 38 patients on dialysis.

- Let y_{ij} be the observed times and δ_{ij} be the indicator ($\delta_{ij} = 1$, if event for patient i, recurrence j occurred; $\delta_{ij} = 0$, right censored) for $i = 1, \cdots, 38$ and $j = 1, 2$.
- With the Proportional Hazards (PH) model, the likelihood contribution of the ith patient in the jth recurrence is (conditional on w_i):
  $L_{ij} = \{h_0(y_{ij}) w_i \exp(\mu_{ij})\}^{\delta_{ij}} \exp\{-H_0(y_{ij}) w_i \exp(\mu_{ij})\}$.
- Assume the baseline hazard to be a Weibull density:
  $h_0(y_{ij}) = \lambda \rho y_{ij}^{\rho-1}$ and $H_0(y_{ij}) = \lambda y_{ij}^{\rho}$.
- Let μ_{ij} be a linear function of 5 covariates; AGE_{ij}, SEX_i and $DISEASE_{ik}$, for $k = 1, 2, 3$ - representing 4 different diseases.
- Finally, the model includes a random effect (frailty term) w_i for patient i.

Example: Testing a Frailty Term.

The null is $Ho : w_i \sim \text{lognormal}(0, \tau)$.

1. Assign "non-informative" priors.
2. To sample from the full conditionals of $\pi(w_i, \beta_h, \rho, \lambda, \sigma^2 | y_{ij}, \delta_{ij})$, the block updating strategies and the Metropolis algorithm within Gibbs are employed.
3. The value S^* in (1) is $S^* = 0.165$.
4. The R package "parfm" is used to estimate βs, ρ, λ, σ^2. Using these estimates, bootstrap samples were generated using the algorithm 7.3 by Davison and Hinkley (1997) and hence the distribution of S.
5. The p-value of S is 0.86 and a lognormal distribution for the frailty is not rejected.