Nonparametric Bayesian literature review

Motivation

- [1] The advantage of using a normal distribution is that the mean parameter plays an important role, and much of the focus is on determining whether or not this mean is 0.

  Typically, this normality assumption is made for the sake of convenience, rather than from some theoretical justification, and may not actually hold.

  For large values of M, the mixture of Dirichlet processes puts most of its mass in narrow tubes, while for small values of M the prior is more diffuse.

- [6] For convenience, the distribution of the random effects is usually taken to be normal, as stated above. Large-sample classical (frequentist) inference regarding β (fixed effects) should not be much affected asymptotically by changing the distribution of the random effects since the first two moments of the marginal distribution of the outcome variable do not depend on the **normality of the distribution** of the random effects. To see this, notice that, marginally,

$$y_i|\beta, \sigma^2, D \sim N_{n_i}(X_i\beta, Z_i D Z_i^T + \sigma^2 I_{n_i}).$$

  For inference regarding β from above, the individual regression coefficients simply result in a complex covariance structure. Bayesian inference for β, using the marginal likelihood implied by above formula, will depend only on (y, σ^2, D). But **the nature of this dependence** will be sensitive to the distributional form ascribed to the b_i.

  The estimated random effect for each individual will be changed by changing the distribution of the random effects. This point is important because there are many applications in which an estimate of the random effect itself is desired. For example, in Tsiatis, DeGruttola, and Wulfsohn (1995), the estimated random effects are used alternately as covariates themselves in a Cox regression model or to create values for time-varying covariates in such a model. ... In such applications, unbiased estimation of the random effects is crucial and the assumption of normality may introduce bias.

  Allowing distributions other than the normal for the random effects may more accurately model our prior beliefs, or it may allow us to better express our uncertainty about the true distribution of the random effects. It is also important to accurately model the distribution of the random effects when prediction for a future observation from a given subject is desired.

  Another situation in which it would be desirable to relax the assumption of normality is when inference is to be made about the distribution of the random effects itself. For example, in randomized complete block designs, the distribution of the random block effects is usually of inferential interest (see John, 1971).

  Another attraction of our approach is that it allows exact Bayesian inference, even in small sample sizes.

- [7] A standard method to allow for this second level of variability is to use random effects. For mathematical convenience these are traditionally assumed to be normally distributed [1], although this is not necessarily appropriate in practice. This assumption for random effects has been criticized in case studies for providing restricted and potentially inappropriate inferences [2], for causing 'over

shrinkage' of the random-effects distribution [3], and for being less efficient under systematic departures from normality [4].

In addition, it is not easy to assess the validity of this normality assumption, particularly when there are only a **small number of higher-level units**; normal probability plots for unobserved latent random effects need to be interpreted with caution, and there is no straightforward diagnostic to test for non-normality [6].

Previous studies suggest that **fixed covariate effects are fairly robust to misspecification of the random-effects distribution [7]**, although such misspecification can more seriously affect the estimates of the variance of the random-effects distribution. Previous results have found **inflated variances for fixed and random effects** when normality is assumed for the random effects in heavy- tailed data [8] or if there is skewing present [9].

- [5] Although estimation for the fixed- effects parameters is relatively robust to misspecification of the random effects, for example, by empirical best linear unbiased estimators (EBLUE) obtained from restricted maximum likelihood (REML) estimation (see Butler and Louis 1992 for some empirical evidence and Jiang 1998 for theoretical results), it is becoming widely recognized that inference for random effects can be misleading when normal distributional assumptions do not hold (Verbeke and Lesaffre 1996).

  Moreover, even though certain features for random-effects distributions, such as variances, can be estimated accurately even when the assumption of normality is violated (Richardson and Welsh 1994; Jiang 1996), detecting higher-order properties of distributions such as skewness and multimodality necessitates dispensing with normality assumptions.

  Identifying such deviations from normality can be important, sometimes leading to critical insight into the data. For example, Zhang and Davidian (2002) used a semiparametric approach to identify population differences in cholesterol levels by detecting skewness in random intercept terms, while Greene (2001) was able to identify clinical differences in the progression of renal kidney disease by identifying skewed and thick-tailed random-slope distributions.

- [3] ... a key to the success of nonparametric Bayes methods in applications is the incorporation of a **sparseness-favoring structure**, which combats the curse of dimensionality.

  It is well known that borrowing of information across subjects is quite sensitive to departures from this assumption. In particular, the normal distribution has light tails and does not allow some subjects to be very different from other subjects or to have groups of subjects that cluster close together. Hence, outlying subjects tend to have their means over-shrunk towards the population mean, and the data from such subjects may be **overly- influential** in estimation of μ.

  Although one could potentially choose a heavier-tailed alternative to the normal random effects distribution, such as a t distribution, it is appealing to instead use a more flexible form that allows for the possibility of skewness and multimodality.

  Even a heavy-tailed distribution, such as the t, has a very restrictive unimodal and symmetric shape, and in most applications there is no reason a priori to believe that such a shape is required.

  Parametric models, such as extensions of the t distribution that allow skewness, are still restrictive, and **do not allow multi-modality**, which may arise due to latent sub-populations. [7]

  ...by assuming that each of the cluster-specific parameters is drawn from the common distribution P_0, we have avoided the over-fitting problem characteristic of assigning subjects to many small clusters.

  Bayesian nonparametric models incorporate **infinitely-many** parameters in order to more flexibly represent uncertainty in P.

  From a Bayesian perspective, in the absence of parametric knowledge of P, one should choose a prior for P with support on the set of distributions on the real line, with this prior effectively corresponding to a distribution over distributions.

By choosing a DP prior for P, one allows P to be an unknown distribution, with P_0 corresponding to one's best guess for P a priori and a expressing confidence in this guess.

- [4] A standard assumption is that of normality of the random effects, but this assumption may not always be realistic, and, because the random effects are not observed, it may be diffcult to verify.

  It is widely acknowledged that the assumption of normality of the random effects may not always be appropriate, see Mallet (1986), Davidian and Gallant (1992, 1993), Fattinger et al. (1995). Complicating matters is the fact that the random effects are unobservable latent model components; thus, no straightforward diagnostic to evaluate the validity of the assumption of normality is available, and analysts typically appeal to the methods implemented in the software without formal evidence that the underlying assumption of normality is appropriate.

  The consequences of misspecifying the random effects distribution have been discussed for linear mixed effects models (Butler and Louis, 1992; Verbeke and Lesaffre, 1996, 1997; Muthen and Shedden, 1999; Tao et al., 1999), and simulations have shown that estimation of fixed parameters may not be severely compromised (e.g. Verbeke and Lesaffre, 1997).

- [11] Often,the normality assumption for the random effects and the error structure is almost automatically taken for granted, and little attention has been devoid to the question of what impact this assumption has on the estimation of the different parts of the model.

  In practice, the estimated r.e. are frequently used to highlight special profiles or to look for (groups of) individuals evolving differently in time. Thus it is very important to investigate to what extent the estimated r.e. distribution represents the true distribution.

- [10] The normality assumption is mathematically convenient and can be robust for estimating the fixed effects(Neuhaus and Hauck, 1992). However, estimation efficiency can be compromised when the random effects distribution is incorrectly specified. In many situations, it is of inherent interest to visualize the distribution or its density, e.g., to ascertain skewness in the outcome profile across individuals.

- [2] The non-parametric and Gaussian approaches produce essentially equivalent inferences for the fixed effects even when the Gaussian assumption for the random effect is violated. This constancy is somewhat surprising in that for the non-parametric approach estimates for a non-linear function of the data.

- [12] Maximum likelihood estimators for fixed effects and variance components in linear mixed models, obtained under the assumption of normally distributed random effects, are shown to be consistent and asymptotically normally distributed, even when the random-effects distribution is not normal.

  Using simulations and the analysis of a real dataset, Butler and Louis (1992) have recently shown that wrongly specifying the prior distribution of univariate random effects has little effect on the fixed effects estimates.

  In Verbeke and Lesaffre (to appear), two datasets were analyzed with a linear mixed-effects model with random effects drawn from a mixture of g multivariate normal distributions. It turned out that varying g had practically no effect on the fixed effects estimates, but also that the residual variance and the covariance matrix of the random effects were very stable. This suggested that the assumption about the random-effects distribution hardly influences any of the parameter estimates.

  Theorem 1. Under general regularity conditions, we have that $\hat{\alpha}_N, \hat{\sigma}_N^2, \hat{D}_N$ are strongly consistent estimators for $\alpha^*, \sigma^{*2}, D^*$, when $N \to \infty$.

  The importance of the above theorem is that it assures that maximum likelihood estimation, based on the assumption of normality for the random effects b_i, yields consistent estimators for all parameters in the model, including the random-effects covariance matrix, even when the true distribution of the b_i is not normal.

Prediction of random effects

- [8] We show that, although the predicted values can vary with the assumed distribution, the prediction accuracy, as measured by mean square error, is little affected for mild-to-moderate violations of the assumptions.

  Our main message is that, although predictions themselves can be sensitive to the assumed distribution, the overall accuracy of prediction is little affected for mild-to-moderate violations of the assumptions.

  ...although the shape of the distribution of the BPs (best predictor) does not necessarily match the shape of the true distribution, this does not necessarily translate into poorer performance in the metric by which BPs are defined, namely mean square error of prediction.

  However, all the random effects models have approximately the same pre- diction error, despite the fact that (Figure 5) the distribution of the BPs from the models are very different.

  mean square error of prediction performance was very robust to the assumed distribution when the random effects variance was small to moderate and cluster sizes were small to moderate. However, for larger variances and larger cluster sizes loss of efficiency can result.

  Very different distributions for BPs can perform quite similarly in practice (as gauged by overall mean square error of prediction).

  Random effects distributions that may be statistically significantly better fitting may not perform better in overall prediction.

  standard approach of assuming Gaussian distributed random effects results in good performance of BP values across a wide range of situations with different true random effect distributions.

- [13] The random terms may be included in the prediction, but averaged over, giving a prediction specific to, or conditional on, the blocks and plots used in the experiment. Alternatively, the predictions may be evaluated at the population marginal mean value (zero) for the random effects. We call these two types of prediction *conditional* and *marginal* for the random effects respectively. Conditional predictions for a factor are appropriate to inference for the specific instance that occurred in the dataset. Marginal predictions for a factor are appropriate when inference is required for members of the wider population.

  Where random effects are part of the predictive model, rather than used purely as error terms, the application and purpose of prediction are important in determining whether conditional or marginal predictions are required, as in the next example.

- [9] When the model parameters are treated as known, the problem of assigning values to random effects can be approached from at least four different philosophical perspectives which we refer to as *Bayesian, empirical Bayesian, frequentist prediction* and *frequentist estimation*. In the **Bayesian approach**, inference regarding ζ_j for cluster j is based on the posterior distribution of ζ_j given the known data for the cluster which are treated as observed values of random variables. **Empirical Bayesians** evaluate inferences with respect to joint sampling of ζ_j and y_j. Robinson (1991) pointed out that this sampling model is also relevant for classical (i.e. frequentist) inference if the problem is viewed as assigning a value to the realization of a random variable. In this case, the random-effects distribution is viewed as representing the variation of ζ_j (in the population), whereas Bayesians would view this prior distribution as representing uncertainty regarding ζ_j.

- (In empirical Bayesian) the parameters are treated as known and equal to their estimates, so the posterior distribution is 'empirical' or 'estimated'. In a fully Bayesian approach, prior distributions would be specified for the model parameters, and the posterior distribution of the random effects would be marginal with respect to these parameters. It should be noted that the estimated posterior distribution can also be derived from a frequentist perspective by treating ?j as unobservable random variables and conditioning on the observed responses y_j (as well as X_j and Z_j).

Miscellaneous

- Because random effects α_i is a random vector rather than a fixed parameter, we talk about predicting α_i rather than estimating α_i.

- Despite the centering of b_i at zero a priori, i.e., $E(b_i|D) = 0$, the random distribution G has a nonzero mean almost surely. This will lead to biased inference for the fixed effects that are paired with the random effects bi, i.e., effects associated with common columns in the design matrices X and Z, e.g., an intercept and a random intercept (Li, Muller, and Lin, 2007).

References

[1] Deborah Burr and Hani Doss. A bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100(469), 2005.

[2] Steven M Butler and Thomas A Louis. Random effects models with non-parametric priors. *Statistics in medicine*, 11(14-15):1981–2000, 1992.

[3] David B Dunson. Nonparametric bayes applications to biostatistics. *Bayesian nonparametrics*, 28:223, 2010.

[4] Alan Hartford and Marie Davidian. Consequences of misspecifying assumptions in nonlinear mixed effects models. *Computational statistics & data analysis*, 34(2):139–164, 2000.

[5] Hemant Ishwaran and Glen Takahara. Independent and identically distributed monte carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association*, 97(460):1154–1166, 2002.

[6] Ken P Kleinman and Joseph G Ibrahim. A semiparametric bayesian approach to the random effects model. *Biometrics*, pages 921–938, 1998.

[7] Katherine J Lee and Simon G Thompson. Flexible parametric models for random-effects distributions. *Statistics in medicine*, 27(3):418–434, 2008.

[8] Charles E McCulloch and John M Neuhaus. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67(1):270–279, 2011.

[9] Anders Skrondal and Sophia Rabe-Hesketh. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):659–687, 2009.

[10] Huageng Tao, Mari Palta, Brian S Yandell, and Michael A Newton. An estimation method for the semiparametric mixed effects model. *Biometrics*, 55(1):102–110, 1999.

[11] Geert Verbeke and Emmanuel Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

[12] Geert Verbeke and Emmanuel Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556, 1997.

[13] Sue Welham, Brian Cullis, Beverley Gogel, Arthur Gilmour, and Robin Thompson. Prediction in linear mixed models. *Australian & New Zealand Journal of Statistics*, 46(3):325–347, 2004.