Test (2011) 20: 65–71
DOI 10.1007/s11749-010-0230-2

DISCUSSION

Rejoinder on: Nonparametric inference based on panel count data

Xingqiu Zhao · N. Balakrishnan · Jianguo Sun

Received: 11 November 2010 / Accepted: 20 November 2010 / Published online: 25 December 2010

We thank the Test editors for organizing the discussion of our article and all of the discussants for their insightful and stimulating contributions. We will provide our response to each of the discussants one by one.

1 Response to Dr. Dean

Dr. Dean touched on the important topic of the efficiency of the nonparametric estimates of the mean function of a point process and the nonparametric comparison procedures discussed in the article. As observed and correctly pointed out by Dr. Dean, the isotonic regression estimator and the test procedures based on the estimator could be less efficient and the reasons for this include the existence of the

Communicated by Domingo Morales.

This rejoinder refers to the comments available at: doi:10.1007/s11749-010-0224-0, doi:10.1007/s11749-010-0225-z, doi:10.1007/s11749-010-0226-y, doi:10.1007/s11749-010-0227-x, doi:10.1007/s11749-010-0228-9, doi:10.1007/s11749-010-0229-8.

X. Zhao
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong
e-mail: xingqiu.zhao@polyu.edu.hk

N. Balakrishnan
Department of Mathematics & Statistics, McMaster University, Hamilton, Canada L8S 4K1
e-mail: bala@mcmaster.ca

J. Sun (✉)
Department of Statistics, University of Missouri, Columbia, MO 65211, USA
e-mail: sunj@missouri.edu

within-subject dependency of recurrent events and the possible complicated forms of the dependency. To overcome this, one natural approach is to specify or postulate the form of such dependency and in doing this, it is obvious that one could misspecify the form. In other words, one would face a trade-off between the efficiency loss and the misspecification effect. In the case of panel count data, it is often difficult to assess both this trade-off and the assumed model for the dependency. It is well known that the efficiency or power of a test procedure depends on many factors including the null and alternative hypotheses, the structure of the underlying response variables or processes involved, the form of the observed data, and the information provided by the observed data. For example, consider the form of the observed data and assume that one is interested in the comparison of several counting processes. In the case of recurrent event data, two common approaches are to formulate the hypothesis by using the intensity process and the rate function, respectively. The former clearly relies on more assumptions and usually yields more efficient test procedures. In practice, however, the latter seems to be preferred more often because of its less assumptions and simplicity.

In the case of panel count data as considered in this article, although one could still model or specify the intensity process or the form of the dependency from the theoretical point of view, one would face the same problems as those in the case of recurrent event data such as model misspecification. A more important and practical issue is that the amount of relevant information contained in panel count data is much less and this would make both the estimation and the model checking more much difficult or even impossible. Based on our experience, we believe that if there does not exist prior information about the dependency and/or the possible structure of underlying point processes, the simple test procedures discussed in the article usually work reasonably well although they are not optimal. However, one needs to be careful if the purpose is to achieve a certain power as in, for example, the design of clinical studies. In these situations, there often exists some prior information that should be considered.

As can be seen from the simulation results presented in the article and discussed by Dr. Dean, the selection of a weight function can have significant effect on the power of a test procedure. When the structures of underlying point processes of interest are simple, one may identify some weight functions that are uniformly better than others. On the other hand, there may not exist a weight function that performs uniformly better. This indicates that one may want to understand the structures first if possible. Dr. Dean suggested that another way for selecting weight functions is to develop a class of weight functions and apply them all together. A similar idea that has been used in practice is to apply all available test procedures to a given comparison problem. It is apparent that different procedures could give completely different results and without any prior information, one could have no way to judge and make the conclusions.

For the analysis of panel count data, unlike the analysis of recurrent event data, one faces two point processes, the underlying recurrent event process of interest and the observation process. If the two processes are not related, the analysis can be performed conditional on the latter process. On the other hand, one has to deal with both and for this, one could apply either the rate function frailty model or the mean function frailty model discussed in the article. It should be noted that they may be

different in general, but under the Poisson assumption, the two models are equivalent. In addition to these two models, one may apply or develop other models such as copula models commented by Dr. Dean. Only limited literature exists on this topic and more research is definitely needed. The same applies to regression analysis of panel count data and for this, as pointed out by Dr. Dean, most of the existing work is based on the proportional intensity or mean models. To follow the ideas commonly used in survival analysis, one could develop models such as additive rates models, additive-accelerated rate models or transformation models (Schaubel et al. 2006; Zeng and Cai 2009; Zeng and Lin 2007).

2 Response to Dr. He

Dr. He discussed the variable selection problem in the context of regression analysis of panel count data. Although our article only provided a brief discussion of the regression analysis and gave no discussion of the variable selection problem, both are clearly important topics in the field of panel count data. Many authors have investigated regression analysis of panel count data, which one often faces in many fields including disease follow-up studies, reliability studies and tumorgenecity experiments. For the variable selection problem with respect to panel count data, as pointed out by Dr. He, however, there basically exists no research in the literature except Tong et al. (2009). The key idea behind the approach proposed by Tong et al. (2009) is to apply the non-concave penalized likelihood approach given in Fan and Li (2001) to the estimating function $W(\beta)$ developed by Sun and Wei (2000) by using Dr. He's notation. We remark that one can replace $W(\beta)$ by a different unbiased estimating function and develop a similar variable selection procedure. Also we want to add that many penalty functions could be used in their approach including the L_p-penalty function (Tibshirani 1996, 1997), the hard thresholding penalty function (Fan 1997) and the smoothly clipped absolute deviation penalty function (Fan and Li 2001).

One area in variable selection that has recently attracted a great deal of attention is high-dimensional data, especially gene expression or microarray studies. For this, many procedures have been developed including dimension reduction approaches and penalized likelihood methods. In these studies, covariates or genes are usually highly correlated and the response variable can be some simple disease markers or patient survival times that may suffer censoring. If the response is a recurrent event process that is observed only at discrete time points, then we would have panel count data with gene expressions as high-dimensional covariates and it is clear that one needs to apply or develop some variable selection procedures. Although many existing procedures could be directly applied to panel count data in theory, one needs to be careful about the special structure of panel count data and without taking into account the structure, the analysis would clearly be inefficient or even invalid.

3 Response to Dr. Ishwaran

We thank Dr. Ishwaran for his thoughtful and incisive comments. Indeed, the weighted gamma process is closely connected to the Dirichlet process and it is therefore natural to exploit the many different methods developed for the Dirichlet process

to carry out the posterior calculations under the weighted gamma process, as Dr. Ishwaran aptly noted. We agree with his point that the weighted gamma process has not been used as much in Bayesian nonparametrics as the Dirichlet process, and it therefore offers much potential for further developments in this regard.

Our likelihood function $L(\mu)$ considered in Sect. 9 is precisely of the form of his likelihood function in (4) wherein

$$f(v) = \sum_{i=1}^{n} \int_0^\infty Y_i(t) F(dt|v)$$

and

$$\prod_{i=1}^{n} \prod_{j=1}^{K_i} \prod_{l=1}^{m_{i,j}} \int_S F(A_{i,j}|v_{ijl}) \mu(dv_{ijl})$$

in place of $\prod_{i=1}^{n} \int_S k_i(X_i, v_i) \mu(dv_i)$. So, the mean of the event process corresponding to a panel count data given by

$$g(\mu) = \int_S F(t|v) \mu(dv)$$

can be estimated, for different values of t, by adopting the three-step Gibbs sampling algorithm that Dr. Ishwaran has outlined. Indeed, the approximations for MCMC algorithms as discussed by Ishwaran and James (2004) could become very useful in this context.

In the case of the analysis of panel count data with covariates, however, the Bayesian computation would become involved. Specifically, in the case of proportional event process model that we have described in Sect. 8, though the mean of the event process could be modeled by

$$\Lambda(t|\mu, \theta) = \exp(\theta' z) \int_S F(t|v) \mu(dv),$$

where $\theta \in R^p$ is the regression parameter and z is a p-dimensional covariate, the normalizing constant C will not be free of the regression parameter θ and so will not cancel out. This will pose some difficulty in the ensuing Bayesian analysis, and it will therefore be of great interest to look further into this problem involving covariates and develop the necessary posterior calculations.

4 Response to Dr. Pardo

Dr. Pardo gave two suggestions and raised two issues that were not touched on in the article. With respect to the nonparametric comparison procedures discussed in the article, Dr. Pardo suggested to replace the estimates of mean functions used in test statistics by the spline-based estimates proposed in Lu et al. (2007). We agree that the new test procedure may give better performance and is worth some investigation.

For treatment comparison, the focus of the article has been on testing the equality of several mean functions. As pointed out by Dr. Pardo, sometimes one may be also interested in testing some linear combinations of the mean functions. For this, by using the notation of Dr. Pardo, one could apply the test statistics

$$\sqrt{n}\int_0^\tau W(t)L'\hat{\Lambda}(t)\,dG_n(t),$$

where $G_n(t) = n^{-1}\sum_{i=1}^n I(T_{i,j} \le t)$ as in the article and $W(t)$ is a weight process. Of course, as pointed out in the article for other comparison problems, another approach is to apply some existing regression methods. In this case, one needs to define some covariates such that the hypothesis of interest can be equivalently represented by a hypothesis on the corresponding regression coefficients.

The analysis of panel count data is a relatively new area and there are a number of issues that have not been investigated in the article or literature. Panel count data from multiple state processes and with measurement errors on covariates are definitely two of them and we wish to thank Dr. Pardo for providing a supplementary review on these two topics. As pointed out by Dr. Pardo, although some approaches have been developed, there exists more work that needs to be done. For example, Kang and Lagakos (2006) developed some likelihood-based methods with semi-Markov models but they did not consider the covariate situation. Also Kim (2007) gave a semiparametric estimation approach for regression analysis of panel count data with covariates measured with errors, but the asymptotic properties of the proposed estimator are still unknown.

5 Response to Dr. Tong

Our thanks are due to Dr. Tong for his interesting comments and suggestions. His point that panel count data involve the observation process and censoring time process both of which are informative of the response variable and that all three processes are related is quite correct. As Dr. Tong has aptly pointed out, an important and challenging problem is to take the dependence between the three processes into account, model them jointly and then develop inferential methods for the parameters of interest.

By taking the conditional means of $N_i(t)$ and $H_i(t)$, given X_i, as

$$E\big(N_i(t)|X_i\big) = \mu_0(t)e^{X_i'\beta}$$

and

$$E\big(H_i(t)|X_i\big) = \lambda_0(t)e^{X_i'\gamma},$$

where β and γ are regression parameter vectors and $\mu_0(t)$ and $\lambda_0(t)$ are baseline mean functions, and then to proceed by assuming the two processes to be conditionally independent given the covariates may not be realistic in some situations. As Dr. Tong has mentioned, it would be reasonable for the recurrence and visit processes to

be correlated in practice. For this reason, the frailty model approach or the latent variable approach to link the two processes is natural to consider, even though they may allow only for positive or negative correlation. However, the recent work of Zhao and Tong (2010) based on the conditional mean specification

$$E\big(N_i(t)|X_i, Z_i\big) = \mu_0(t)g(Z_i)e^{X_i'\beta},$$

where $\mu_0(t)$ is the baseline mean function, $g(\cdot)$ is an unspecified function and β is the regression parameter vector, alleviates some of these difficulties. These authors have then used estimating function approach to estimate the parameters of interest. Even though the function $g(\cdot)$ cannot be estimated in this approach, it is a promising new development as it facilitates capturing the correlation between the two processes. Many more problems remain open in this direction. One question is whether the function $g(\cdot)$ could be restricted to be in a specified family and then determine an optimal $g(\cdot)$ within that family. Having said this, the most challenging problem that remains is the joint modeling of all three processes and then developing the corresponding estimation method, as properly stated by Dr. Tong.

6 Response to Dr. Uña-Álvarez

As Dr. Pardo, Dr. Uña-Álvarez also raised the question of analyzing multiple state panel count data and, in particular, pointed out some direction for nonparametric estimation and the possible connection with interval-censored failure time data. It is true that in a recurrent event study, if the event can happen only a small number of times, the study and the data could be described by a multiple state model. When doing this, one has to pay attention to two aspects. One is that in the case of recurrent event data, the difference may not be significant from the analysis point of view, while both the modeling and the analysis tools required may be quite different for panel count data due to the incomplete nature of the data. The other is that the problems to be addressed tend to be different. With respect to the connection to interval-censored data, if the recurrent event process of interest is a one-jump counting process, it is clear that we could regard the process as a survival process and apply some developed techniques for interval-censored data to panel count data. However, for failure time data, the main focus behind the development of statistical tools is to deal with a censoring mechanism, which does not usually exist or may not be the focus in recurrent event studies. In other words, sometimes it may be easier or more convenient to directly develop approaches for panel count data. We agree that it would be interesting to modify the Aalen–Johansen (Aalen and Johansen, 1978) estimator of a transition matrix to deal with multiple state panel count data. As pointed out by Dr. Uña-Álvarez, however, this is indeed very difficult as we only know the states of each study subject at finite distinct observation times and the exact jumping times of the underlying process are not available.

As discussed by Ghosh and Lin (2002) and others, for recurrent event data, a dependent terminal process could exist and make the analysis more difficult. In the case of panel count data, in addition to the terminal process, there also exists an

observation process that does not exist for the former case. Of course, one may perform the analysis of panel count data conditional on the observation process if the process is noninformative. Otherwise, one has to model it and in this case, note that we have recurrent event data and could model it through, for example, a proportional rates model. If we further assume a proportional means model for the recurrent event process, then some techniques similar to those used in Ghosh and Lin (2002) can be developed although much more complicated due to the involvement of an extra process. Some detailed discussion on this can be found in He et al. (2009) and Huang et al. (2006) among others.

We appreciate very much that Dr. Uña-Álvarez brought to our attention and briefly introduced two R packages for the analysis of panel count data. Unfortunately we have not had any experience with the two packages yet.

References

Aalen O, Johansen S (1978) An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. Scand J Stat 5:141–150

Fan J (1997) Comment on "Wavelets in statistics: a review" by A. Antoniadis. J Ital Stat Soc 131–138

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 1348–1360

Ghosh D, Lin DY (2002) Nonparametric analysis of recurrent events and death. Biometrics 56:554–562

He X, Tong X, Sun J (2009) Semiparametric analysis of panel count data with correlated observation and follow-up times. Lifetime Data Anal 15:177–196

Huang C-Y, Wang M-C, Zhang Y (2006) Analysing panel count data with informative observation times. Biometrika 93:763–775

Ishwaran H, James LF (2004) Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. J Am Stat Assoc 99:175–190

Kang M, Lagakos SW (2006) Statistical methods for panel data from a semi-Markov process, with application to HPV. Biostatistics 8:252–264

Kim J (2007) Analysis of panel count data with measurement errors in the covariates. J Stat Comput Simul 77:109–117

Lu M, Zhang Y, Huang J (2007) Estimation of the mean function with panel count data using monotone polynomial splines. Biometrika 94:705–718

Schaubel DE, Zeng D, Cai J (2006) A semiparametric additive rates model for recurrent event data. Lifetime Data Anal 389–406

Sun J, Wei LJ (2000) Regression analysis of panel count data with covariate-dependent observation and censoring times. J R Stat Soc B 62:293–302

Tibshirani RJ (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 267–288

Tibshirani RJ (1997) The lasso method for variable selection in the Cox model. Stat Med 385–395

Tong X, He X, Sun L, Sun J (2009) Variable selection for panel count data via nonconcave penalized estimating function. Scand J Stat 36:620–635

Zeng D, Cai J (2009) Additive-accelerated rate model for recurrent event. In: New developments in biostatistics and bioinformatics. Front stat, vol 1. World Sci Publ, Singapore, pp 35–48

Zeng D, Lin DY (2007) Semiparametric transformation models with random effects for recurrent events. J Am Stat Assoc 167–180

Zhao X, Tong X (2010) Semiparametric regression analysis of panel count data with informative observation times. Comput Stat Data Anal. doi:10.1016/j.csda.2010.04.020