Comments on: Model-free model-fitting and predictive distributions : Applications to Small Area Statistics and Treatment Effect Estimation

Sperlich, Stefan Andréas

Comments on Model-free model-fitting and predictive distributions: Applications to Small Area Statistics and Treatment Effect Estimation

Stefan Sperlich*

March 5, 2014

Abstract

Discussing the paper "Model-free model-fitting and predictive distributions" by Politis (2013), we propose to extend this procedure to semiparametric and parametric mixed effects models (MEM) as in practice, these are probably the most popular ones for prediction. Specifically, combining Politis' prediction method with procedures from Lombardía and Sperlich (2012) and González-Manteiga, Martínez-Miranda, Lombardía-Cortiña and Sperlich (2013) yields new MEM-based MF/MB point and interval predictors which can be used for example for small area statistics. Combining Politis' idea with nonparametric matching estimators may also yield improved (point and interval) estimators for treatment effects and policy evaluation.

Keywords: prediction, small area statistics, treatment effect estimation, non- and semiparametric estimation.

1 Introductory Comments

The paper of Politis (2013) provides a new and original way of thinking about prediction[1]. The *model free prediction principle* arms him with a universal tool that goes clearly beyond the classical understanding of conditional mean and quantile prediction. Importantly, the provided resampling methods extend this idea to further construct reliable prediction intervals instead of just point predictors. In his Sections 3 and 4 he shows very well that this principle is indeed feasible and can often be reached by slight modifications of well-known standard estimation methods and bootstrap. His simulations provided in

*Département des sciences économiques and Research Center for Statistics, Université de Genève, Bd du Pont d'Arve 40, CH-1211 Genève, Suisse. E-mail address: stefan.sperlich@unige.de. The author acknowledges funding from the Swiss National Science Foundation 100018-140295.

[1] Admittedly, he has already been applying and developing this idea further since a couple of years, as you can see from his own reference list.

different Subsections demonstrate how powerful the resulting predictors are compared to the procedures having been used so far.

It is evident that given the feasibility and excellent performance, this prediction principle will open a new field of research in at least three directions: **(a)** the domain of possible applications, which is the one we will mainly discuss in the following sections, **(b)** the domain of practical issues like the choice of transformation and priors, namely of the model or smoothing parameters, the smoother itself (thinking of kernels, local polynomials, penalized or regression splines, etc.), dimension reduction (semiparametric or additive models, essential dimension reduction, etc.), resampling alternatives (subsampling maybe, see Politis, Romano and Wolf, 1999) and the measurement of predictive power[2], **(c)** and last but not least the domain of efficient and user-friendly implementation.

2 Extension to Mixed Effects Models

When Politis (2013) speaks of 'model-free', may it be in terms of MF/MB or as MF/MF[2], it should not be unnoticed that the model choice, including the nonparametric option will have an impact on the functioning of the prediction principle and therefore also the quality (i.e. size and location) of the prediction interval. A particular obstacle is the potential dependency between the variable of interest, i.e. the responses Y in his notation. While in a model-based approach this dependency can often be handled, even if the model is semi- or nonparametric; overcoming such a problem is more challenging in his model-free approach discussed in his Section 4. The case of time series has been considered by him in different papers, if only in the MF/MB context.

Being aware of both the benefits and withdraws of such a dependency, statisticians in the different fields where such dependencies are obvious or prediction is of major interest, have developed strategies along mixed effects models; for longitudinal data in biometrics, for panel data in econometrics, for area-clustered data in official statistics including environmetrics and poverty mapping (also well-known as 'small area statistics'), or think of the extension to more complex hierarchical models in social sciences. Limiting the discussion to random intercepts, equation (1) of Politis (2013) then extends to a model of the form

$$Y_{d,t} = \mu(\underline{x}_{d,t}) + u_d + \sigma(\underline{x}_{d,t})\varepsilon_{d,t} \quad \text{for } d = 1, \ldots, D, \text{ and } t = 1, \ldots, n_d \tag{1}$$

where the index d indicates to which cluster (area) the observation belongs to, and $\sum_{d=1}^{D} n_d = n$ is the sample size. When referring to the population we use $N_d < \infty$ with $D \to \infty$ asymptotically. While Y, $\underline{x}$, and ε are as in Politis (2013), u_d stands for the area specific effects. If this is independent from ε and $\underline{x}$, it can be taken and treated as a

[2]It might be of interest in this context that the proposal by Nielsen and Sperlich (2003) of such a measure, the so-called validated R^2, is simply the classical R^2 but calculated from the predictive residuals of Politis (2013) in the numerator and denominator. Nonetheless, by parts of the community working on empirical finance it has typically been refused as a measure of predictive power, certainly without any sophisticated justification.

random effect. If some dependency between the area and $\underline{x}$ is likely, then one must account for this by first filtering out the common effect, see Lombardía and Sperlich (2012).

The literature about estimation, inference and prediction with mixed models of the form (1) is abundant, so that we can only mention few recent contributions closely related to our context: Verbeke and Molenberghs (2000) is a standard compendium for linear mixed models for longitudinal data with $\mu(\cdot)$ being a linear model and d indicating the subject with n_d repeated measurements over time. Wand (2003) combines spline smoothing and mixed models where $\mu(\cdot)$ is a series of orthogonal polynomials, d indicates an area of the support of $\underline{x}$, and u_d the local deviation from the polynomials. Lombardía and Sperlich (2008) propose estimators, tests and bootstrap procedures for generalized nonparametric mixed models; $\mu(\cdot)$ is an unknown smooth function of $\underline{x}$, and u_d the random effect. González-Manteiga, Martínez-Miranda, Lombardía-Cortiña and Sperlich (2013) review the various existing nonparametric estimators for model (1).

It is not hard to see that all these models can be embedded in the MF/MB prediction principle of Politis (2013) almost straight forwardly; the linear-model-based methods of Verbeke and Molenberghs (2000) for example in his Sections 3.6 and 3.7, the others mentioned above in 3.1 and 3.5 – they all refer to non- and semiparametric models. It is mainly the bootstrap algorithm, cf. his Sections 2.6 and 3.5, respectively, that has to be adapted as we have to handle two random terms, namely u_d and $\varepsilon_{d,t}$ in our mixed model (1). Furthermore, while the latter ($\varepsilon_{d,t}$) are assumed to be i.i.d., the u_d might have a (conditional) covariance structure $\Sigma_u(\underline{x})$. Following Lombardía and Sperlich (2008), one draws some random $u_d^*|\underline{x}$ with mean zero and covariance $\widehat{\Sigma}_u(\underline{x})$ but can proceed for the $\varepsilon_{d,t}^*$ as in Politis (2013) to obtain the bootstrap sample $(y^*, \underline{x})$ with[3] $y*_{d,t} := m(\underline{x}_{d,t}) + u_d^* + s_{x_{d,t}}\varepsilon_{d,t}^*$. From that sample one finally calculates the bootstrap estimates $m^*(\underline{x}), s_x^*$ and predictors $\hat{u}_d^*$. In other words, the resampling algorithm of Politis (2013) is almost unchanged but enlarged by one step, say *(a.2)* to draw the u_d^*, which are afterward added to Y^* in *(b)*, *(c)* and predicted ($\hat{u}_d^*$) in *(d)*. By these steps we account for the random effect u_d.

When turning to the construction of prediction intervals, notice that the incertitude of the prediction depends, as before, on the variance of the estimator of $\mu(\cdot)$, the individual variation σ_x and its estimation error, but additionally also on the prediction error of u_d. This is automatically taken into account in step *(e)*, namely[4]

$$prediction - error^* = g(Y_{d,f}^*) - \Pi(g, m^*(\underline{x}_{d,f}), s_{x_{d,f}}^*, \hat{u}_d^*, \mathbf{X}, \hat{F}_n^*) \ , \tag{2}$$

where $g(Y_{d,f})$ is the prediction of interest for a known function $g(\cdot)$ and $\underline{x}_{d,f}$ of a subject in a given area (or cluster) d. The predictor $\Pi(\cdot)$ has to be chosen from Table 3.1 of Politis (2013) completed by the corresponding adding[5] of $\hat{u}_d$. As in Politis (2013), $\hat{F}_n^*$

[3] We try to keep here almost the notation of Politis (2013), where $m(\underline{x}) = \widehat{\mu}(\underline{x})$, $s_x = \widehat{\sigma}(\underline{x})$, and $\hat{F}_n$ being the empirical distribution of the residuals.

[4] Again we follow the notation of Politis (2013).

[5] You simply substitute $m(\underline{x}_{d,f}) + \hat{u}_d$ in Table 3.1. for m_{x_f}.

indicates the empirical distribution of the ε^*, and $\mathbf{X}$ the set of all observed (or used) explanatory variables. Consequently, it seems that no further adaptation of Politis' proposal for constructing prediction intervals is necessary.

Unfortunately, in mixed effects models the choice of bandwidth and consequently also the choice of the bootstrap bandwidth becomes even more crucial and complex. This is especially true when some of the explanatory variables vary only over the different areas $d = 1, \ldots, D$. An inappropriate bootstrap bandwidth does not always lead to an undercoverage, it can equally well lead to an overcoverage. In any case it distorts the correctness of the prediction interval, see Remark 3.3 of Politis (2013). Sperlich (2013) explains that this bandwidth problem is similar to that of finding the optimal subsample size in subsampling (see Politis, Romano and Wolf, 1999). Consequently, in order to find an appropriate bootstrap bandwidth he proposes – with some success in his simulations – to apply the same procedure as it is used for determining an appropriate subsample size.

3 Its use in Small Area Statistics

Small area statistics is one of the fields to which mixed effects models owe their extraordinary popularity. Again, the literature about small area statistics is abundant: see for example Rao (2003) for a classical compendium, and Jiang and Lahiri (2006) for a review of more recent contributions. Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008) and Sperlich and Lombardía (2010) were probably the first presenting nonparametric small area estimators and specification tests based on mixed effects models. Salvati, Chadra, Ranalli and Chambers (2010) discussed nonparametric direct estimators for small areas. Lombardía and Sperlich (2012) showed how the necessary independence of the random effects can be reached.

The problem is always to estimate or predict a certain parameter for all small areas, mostly the mean but sometimes certain quantiles or even a special distribution parameter which indicates e.g. the percentage of poor in an area. The available data can be census data with some missing responses Y (typically assumed to be missing at random) or sample data with some missing Y. Let us consider in the following the simple example of predicting the missing responses Y in a census to afterward predict a linear combination (like e.g. the mean) of the Y in each area, say $\gamma_d' \underline{Y}_d$ where $\underline{Y}_d = (Y_{d,1}, \ldots Y_{d,N_d})$ for given $\gamma_d \in I\!R^{N_d}$. Let us split it to $\gamma_d = (\gamma_{d,s}', \gamma_{d,f}')'$ where the first part of the vector stands for the observed Y, and the second for the ones which have to be predicted. Analogously we split $\underline{Y}_d = (\underline{Y}_{d,s}', \underline{Y}_{d,f}')'$ of which $\underline{Y}_{d,f}$ is unobserved, and $\underline{x}_d = (\underline{x}_{d,s}', \underline{x}_{d,f}')'$ which consists only of observed explanatory variables. Holding census data (in this example), the for the prediction interval relevant prediction error comes exclusively from $\gamma_{d,f}' \hat{\underline{Y}}_{d,f}$.

Following exactly the prediction and bootstrap procedure from above, Section 2, we get by resampling as many

$$prediction - error^* = \gamma_{d,f}' \hat{\underline{Y}}^*_{d,f} - \gamma_{d,f}' \{m^*(\hat{\underline{x}}_{d,f}) + \hat{u}^*_d + s^*_x \varepsilon^*_{d,f}\}$$

as we need (or want) to simulate the distribution of the real prediction error. From these one can easily construct now any prediction interval for $\gamma_d' \underline{Y}_d$, and analogously for any other area $d = 1, \ldots, D$.

We must be aware that albeit the excellent properties Politis' procedure exhibits in his simulations should carry over to this problem, the bad properties of predictors not following his procedure might not. The reason is that the prediction problem is somewhat easier as it is just a summary statistic like the average of predictors (i.e. where $\gamma_d = (1, 1, \ldots, 1, 1)/N_d$) that matters. In particular, it has to bee studied whether and to what extend the superiority or discrepancy of his method is maintained or even increases for an increasing number of missing responses Y.

4 Improved Matching Estimators for Average Treatment Effects (of the Treated)?

Treatment effect estimation has a long history in biometrics. During the last fifteen years it was experiencing a revival when it finally was detected by the econometricians. We start with explaining in brief why in first line this is actually a problem of prediction rather than of estimation. To this aim, let T be the treatment indicator.

The interest is to estimate the effect of a treatment ($T = 1$) on 'output' Y versus the counterfactual case of no treatment ($T = 0$) for the same person, and the population average of this effect. The important thing to understand is that we speak of the same person, i.e. treatment effect $TE_i = Y_i^1 - Y_i^0$ of person i or its average $ATE = E[Y_i^1] - E[Y_i^0]$. Often one is already happy with obtaining a good estimate for the average treatment effect for the treated (only)

$$ATET = E[Y_i^1 | T_i = 1] - E[Y_i^0 | T_i = 1] \ . \tag{3}$$

For the ease of notation let us concentrate on the latter. It is clear that $(\sum_{i:Ti=1} Y_i)/(\sum_{i:Ti=1} T_i)$ is a reasonable estimator for $E[Y_i^1 | T_i = 1]$. The real problem is to construct an estimator for $E[Y_i^0 | T_i = 1]$. With a control group of non-treated subjects at hand we are tempted to set

$$\widehat{E}[Y_i^0 | T_i = 1] = (\sum_{i:Ti=0} Y_i)/(\sum_{i:Ti=0} T_i) \ . \tag{4}$$

Unfortunately, this is only a good idea if both groups are random samples of the total population, a rather unrealistic assumption in practice. Therefore, what remains unsolved is the prediction problem (4) to forecast the potential non-treatment outcome Y^0 for the treated with the aid of the non-treated control group.

Formally spoken, in order to use (4) one would need the independence condition $T \perp (Y^1, Y^0)$ which typically is unlikely to hold in social sciences unless we can force people to treatment and exclude them equally well. Having so-called 'confounders' $\underline{x}$ at hand, i.e.

variables that have an impact on T but not vice verse at the time of its observation, one may hope for the conditional independence $T \perp (Y^1, Y^0)|\underline{x}$, the CIA. Given CIA one has

$$E[Y|T=t,\underline{x}] = E[Y^t|T=t,\underline{x}] = E[Y^t|\underline{x}] =: \mu_t(\underline{x}) \quad , t = 0, 1, \tag{5}$$

where $\mu_0(\cdot)$ can be estimated from the observations of the control group. As

$$ATET = \int\ Y - E[Y^0|T=1,\underline{x}]\ dF(\underline{x}|T=1) = \int\ Y - \mu_0(\underline{x})\ dF(\underline{x}|T=1), \tag{6}$$

cf. equation (3), a resulting matching estimator is simply

$$\widehat{ATET} = \sum_{i:T_i=1} \left(Y_i^1 - \sum_{j:T_j=0} W_{i,j} Y_j^0 \right) / (\sum_{i:T_i=1} T_i) \tag{7}$$

with $W_{i,j}$ a weight function regarding closeness of j to i 'in $\underline{x}$' and summing up to 1 (like kernel smoothers do), see Abadie and Imbens (2006) for more details and discussion.

In other words, as already mentioned above and can be seen now from (7), one forecasts Y^0 for subject i in the treatment group by matching it with members of the control group which are similar to i in their characteristics. The error distribution of these matching estimators is typically approximated by bootstrap methods[6]. We conclude this section with basically the same remark that we mentioned at the end of Section 3. The number of needed predictions is large, if not huge, and can easily amount to several thousands to be integrated over afterwards. Consequently, it would be interesting to see whether the superiority of Politis' method is then still maintained, increases or vanishes.

5 Concluding Remarks

It would be too lengthy to discuss the further thinkable, often quite straight extensions of what is outlined above. In Section 2 one might continue with multilevel models including some more nested random effects, discuss the particular implementation via prior modes or (pseudo) likelihood functions, or switch to generalized (linear) mixed effects models with known link function.

In Section 3, apart from the just mentioned ones, it is most interesting to see how the method carries over to problems where specific distribution parameters are demanded that are more complex than the mean or a quantile; think of the various definitions of a poverty line and how to estimate the proportion(s) of the area-population(s) lying below it, cf. Elbers, Lanjouw and Lanjouw (2003).

In Section 4 the first question coming up to my mind is the extension to propensity score based (or propensity score weighted) matching. There, the full information $\underline{x}$ gets reduced

[6] However, to the best of my knowledge, the consistency of bootstrap for nonparametric matching estimators has so far not been proved in general. In contrast, Abadie and Imbens (2008) have proved the inconsistency of the bootstrap procedure for the often used kNN matching.

to the so-called 'propensity to treatment' $P(T = 1|\underline{x})$ as this is supposed to contain all relevant information, functioning thus like a sufficient statistic. The obvious advantage is the dimension reduction; in econometrics $\underline{x}$ often amounts to a set of about twenty to thirty variables whereas the propensity score regression is just one dimensional. Further desirable extensions comprise the ATE estimation or the difference-in-difference matching.

All these extensions, however, are more or less straight (even though not always trivial). Personally, I am curious about the seemingly more challenging question how to extend the MF/MF^2 idea of Section 4 of Politis (2013) to those kinds of models and prediction problems. Certainly, one could try to work in e.g. the random effects or the (treatment) selection process in Politis' procedure. But when doing so we are easily tempted to switch to a model-based approach. On the other hand we could equally well ask whether a MF/MF^2 is at all desirable if we already have knowledge about some structure in the data, and how to decide which approach is superior then? We could even go a step further and ask the same question that has been bothering statisticians since decades if not centuries: the one of model selection, see e.g. Linhart and Zucchini (1986) and Claeskens and Hjort (2008), or (if using nonparametric models) bandwidth selection, cf. Köhler, Schindler and Sperlich (2013).

References

Abadie, A. and Imbens,G. (2006) Large Sample Properties of Matching Estimators for Average Treatment Effects, *Econometrica*, **74**, 235-267.

Abadie, A. and Imbens, G. (2008) On the failure of the bootstrap for matching estimators, *Econometrica*, **76**, 15371557.

Claeskens, G. and Hjort, N.L. (2008) *Model Selection and Model Averaging*, Cambridge University Press.

Elbers, C., Lanjouw J.O. and Lanjouw P. (2003) Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355-364.

González-Manteiga, W., Martínez-Miranda, M.D., Lombardía-Cortiña, M.J. and Sperlich, S. (2013) Kernel Smoothers and Bootstrapping for Semiparametric Mixed Effects Models, *Journal of Multivariate Analysis*, **114**, 288-302.

Jiang, J. and Lahiri, P. (2006) Mixed Model Prediction and Small Area Estimation. *Test*, **15**, 1-96, with discussion.

Köhler, M. Schindler, A. and Sperlich, S. (2013) A Review and Comparison of Bandwidth Selection Methods for Kernel Regression. *International Statistical Review*, conditionally accepted.

Linhart, H. and Zucchini, W. (1986) *Model Selection*, Wiley, New-York.

Lombardía, M.J. and Sperlich, S. (2008) Semiparametric Inference in Generalized Mixed Effects Models, *Journal of the Royal Statistical Society B*, **70**, 913-930.

Lombardía, M.J. and Sperlich, S. (2012) A new class of Semi-Mixed Effects Models and its Application in Small Area Estimation, *Computational Statistics & Data Analysis*, **56**, 2903-2917.

Nielsen, J.P. and Sperlich, S. (2003) Prediction of stocks: A new way to look at it, *Astin Bulletin*, **33**, 399-417.

Opsomer, J., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2008) Nonparametric Small Area Estimation Using Penalized Spline Regression, *Journal of the Royal Statistical Society, Series B*, **70**, 265-286.

Politis, D.N. (2013) Model-free model-fitting and predictive distributions, *Test*, this issue.

Politis, D.N., Romano, J.P. and Wolf, M. (1999) *Subsampling*, Springer, New-York.

Rao, J.N.K. (2003) *Small Area Estimation*, John Wiley and Sons, Inc., New-York.

Salvati, N., Chadra, H., Ranalli, M.G. and Chambers, R. (2010) Small area estimation using a nonparametric model-based direct estimator, Computational Statistics & Data Analysis, **54**, 2159-2171.

Sperlich, S. (2013) On the Choice of Regularization Parameters in Specification Testing: A critical discussion, *Empirical Economics*, conditionally accepted.

Sperlich, S. and Lombardía, M.J. (2010) Local Polynomial Inference for Small Area Statistics: Estimation, Validation and Prediction, *Journal of Nonparametric Statistics*, **22**, 633-648.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*, Springer, New-York.

Wand, M.P. (2003) Smoothing and Mixed Models, *Computational Statistics*, **18**, 223-249.