TI 2017-029/III
Tinbergen Institute Discussion Paper

Weighted-Average Least Squares Estimation of Generalized Linear Models

Giuseppe de Luca[1]
Jan R. Magnus[2]
Franco Peracchi[3]

[1] *University of Palermo, Italy*
[2] *VU Amsterdam and Tinbergen Institute, The Netherlands*
[3] *University of Rome Tor Vergata and EIEF, Italy*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Weighted-average least squares estimation of generalized linear models[*]

Giuseppe De Luca

University of Palermo, Italy

Jan R. Magnus

Vrije Universiteit Amsterdam and Tinbergen Institute, The Netherlands

Franco Peracchi

University of Rome Tor Vergata and EIEF, Italy

February 25, 2017

Abstract

The weighted-average least squares (WALS) approach, introduced by Magnus et al. (2010) in the context of Gaussian linear models, has been shown to enjoy important advantages over other strictly Bayesian and strictly frequentist model averaging estimators when accounting for problems of uncertainty in the choice of the regressors. In this paper we extend the WALS approach to deal with uncertainty about the specification of the linear predictor in the wider class of generalized linear models (GLMs). We study the large-sample properties of the WALS estimator for GLMs under a local misspecification framework that allows the development of asymptotic model averaging theory. We also investigate the finite sample properties of this estimator by a Monte Carlo experiment whose design is based on the real empirical analysis of attrition in the first two waves of the Survey of Health, Ageing and Retirement in Europe (SHARE).

Keywords: WALS, model averaging, generalized linear models, Monte Carlo, attrition

JEL classification: C51; C25; C13; C11

[*] Corresponding author: Giuseppe De Luca (`giuseppe.deluca@unipa.it`). We thank Gerda Claeskens and Henk Pijls for useful comments and discussions.

1 Introduction

In recent years, a large body of the statistics and econometrics literature has been concerned with the development of inferential methods to address a variety of model uncertainty problems. The two most popular approaches are model selection and model averaging. In model selection, the investigator first chooses a best performing model according to some criterion and then carries out inference based on the chosen model by ignoring the uncertainty due to the initial model selection step. This popular approach is subject to many problems, but the most important is that the model selection step is completely separated from the estimation step. As shown by Magnus (1999, 2002), Leeb and Pötscher (2003, 2006), and Berk et al. (2013), among others, the initial model selection step matters and is likely to have nonnegligible effects on the statistical properties of the resulting estimates.

Model averaging, on the other hand, provides a more satisfactory approach to inference because it does not require the investigator to rely on a single 'best' performing model. Based on the idea that each model contributes information on the parameters of interest, one computes a weighted average of the conditional estimates across all possible models to combine the available pieces of information into an unconditional estimate that incorporates the uncertainty due to both the model selection and the model estimation steps. A distinction can be made between four types of model averaging methods depending on whether the estimation of each model and the choice of the associated weighting scheme are developed along frequentist or Bayesian lines. These different methods have originated a rapidly expanding literature on model averaging, including in particular a variety of strictly Bayesian (BMA) and strictly frequentist (FMA) model averaging estimators. Useful overviews of the two approaches can be found in Hoeting et al. (1999), Clyde and George (2004), Claeskens and Hjort (2008), and Moral-Benito (2015).

Model averaging is not the only way to allow for uncertainty due to both model selection and estimation, and shrinkage and penalized methods are also receiving increasing attention. Recent work by Hansen (2014, 2016) shows that Stein-type shrinkage estimators can be interpreted as model averaging estimators in the case of two nested models. Methods that simultaneously select variables and shrink coefficients by minimizing some penalized loss function include, among others, the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996) and the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001). Bayesian counterparts of these frequentist approaches are also available. For example, the Bayesian LASSO of Park and Casella (2008) is motivated by the fact that the LASSO estimate of linear regression parameters can be interpreted as a posterior mode when the regression parameters have independent Laplace priors. Further, as noticed by Kumar and Magnus (2013), the LASSO and SCAD estimators can be interpreted as discontinuous counterparts of the Laplace, Subbotin and reflected Weibull estimators available in a Bayesian context. LASSO-type methods have been shown to be particularly effective in high-dimensional settings where the number of predictors exceeds the sample size (see, e.g., Fan and Lv 2010, Chernozhukov et al. 2015, and Belloni et al. 2017), but recent work by Ando and Li (2014, 2017) suggests that model averaging procedures also perform well in these more complex settings.

In this paper we focus on the weighted-average least squares (WALS) approach introduced by Magnus et al. (2010) to account for model uncertainty in the choice of the regressors for a Gaussian linear model. The WALS estimator is a Bayesian combination of frequentist estimators: the parameters of each model are estimated by least squares under a classical frequentist perspective, while the weighting scheme is based on a Bayesian perspective using posterior model probabilities to reflect the confidence in each model based on prior beliefs and the observed data. The result of this 'Bayesian-frequentist fusion' is a model averaging estimator that has some important advantages over standard BMA and FMA estimators. First, in contrast to several BMA estimators that adopt normal priors leading to unbounded risk, the choice of prior in WALS is based on theoretical considerations related to admissibility, bounded risk, robustness, near-optimality in terms of minimax regret, and proper treatment of ignorance (see, e.g., Magnus 2002, Magnus et al. 2010, Kumar and Magnus 2013, and Magnus and De Luca 2016). Second, unlike BMA and FMA estimators, WALS uses a preliminary semiorthogonal transformation of the regressors that allows to obtain exact model-averaging estimates of the parameters of interest in negligible computing time.

The aim of this paper is to extend the WALS approach to deal with uncertainty about the specification of the linear predictor in the wider class of generalized linear models (GLMs). This class includes a variety of nonlinear models for discrete and categorical outcomes, such as logit, probit, and Poisson regression models. A previous attempt to extend the WALS methodology in the same direction was undertaken by Heumann and Grenke (2010), but their paper is restricted to the logit model and lacks a rigorous treatment of the underlying theory. Our paper provides a more comprehensive treatment of the WALS approach to GLMs and establishes the large-sample properties of this class of model averaging estimators under the local misspecification framework proposed by Hjort and Claeskens (2003a).

Specifically, we show that many of the theoretical and computational advantages of the WALS approach to Gaussian linear models continue to hold in the wider class of GLMs by a simple linearization of the constrained maximum likelihood (ML) estimators. To establish the asymptotic theory for WALS, some improvements had to be made to the semiorthogonal transformation procedure. These improvements address potential discontinuity problems in the eigenvalue decomposition used in earlier papers on WALS. In addition to developing the asymptotic theory for the WALS estimator of GLMs, we also investigate the finite-sample properties of our model averaging estimator by a Monte Carlo experiment whose design is based on a real empirical example, namely the analysis of attrition in the first two waves of the Survey of Health, Ageing and Retirement in Europe (SHARE). Here, we compare the performance of WALS with that of other popular estimation methods such as standard ML, strict BMA with conjugate priors for GLM (Chen and Ibrahim 2003; Chen et al. 2008), and strict FMA with weighting systems based on smooth information criteria (Buckland et al. 1997; Hjort and Claeskens 2003a).

The remainder of the paper is organized as follows. Section 2 presents the statistical framework. Section 3 discusses some properties of ML estimators that are important for constructing WALS estimators of GLMs. Section 4 discusses WALS estimation. Section 5 presents an empirical illustration. Section 6 presents a set of Monte Carlo simulations. Finally, Section 7 concludes.

2 Statistical framework

We consider modeling a data matrix $[y : X]$ consisting of n observations on a scalar outcome and k regressors. Thus, y is an n-vector with ith element y_i and X is an $n \times k$ matrix with ith row x_i'. As in a standard GLM setup, we assume that the elements of y are realizations of n independently distributed random variables with mean μ_i, finite nonzero variance σ_i^2, and distribution belonging to the one-parameter linear exponential family (LEF) with density (or probability mass function)

$$f(y_i;\theta_i) = \exp\left[\,\theta_i\, y_i - b(\theta_i) + c(y_i)\,\right], \tag{1}$$

where θ_i is a scalar parameter called the canonical parameter, $b(\cdot)$ is a known, strictly convex and twice continuously differentiable function, and $c(\cdot)$ is a known function. Different choices of $b(\cdot)$ and $c(\cdot)$ result in different distributions within the LEF (e.g., normal, binomial or Poisson). In the original formulation of Nelder and Wedderburn (1972), the density of y_i also includes a dispersion parameter which, without loss of generality, we set equal to one. By the properties of the LEF, the mean and variance of y_i are equal to $\mu_i = \mu(\theta_i)$ and $\sigma_i^2 = \sigma^2(\theta_i)$, with $\mu(\theta) = db(\theta)/d\theta$ and $\sigma^2(\theta) = d^2b(\theta)/d\theta^2 = d\mu(\theta)/d\theta$ (McCullagh and Nelder 1989). The assumptions on $b(\cdot)$ guarantee that the function $\mu(\cdot)$ is invertible and the function $\sigma^2(\cdot)$ is strictly positive.

As in a standard GLM setup, we model the dependence of y_i on x_i by assuming that there exist a linear predictor $\eta_i(\beta) = x_i'\beta$ and an invertible and continuously differentiable function $h(\cdot)$, called the inverse link, such that

$$\mu(\theta_i) = \mu_i = h(\eta_i(\beta)) \tag{2}$$

for a unique point β in a k-dimensional parameter space. When $h(\cdot) = \mu(\cdot)$ (the 'canonical link case'), this assumption corresponds to a linear model $\theta_i = x_i'\beta$ for the canonical parameter. More generally, assumption (2) implies that the canonical parameter θ_i is a smooth function of the linear predictor η_i, written $\theta_i = \theta(\eta_i)$ where $\theta(\cdot) = \mu^{-1}(h(\cdot))$.

We assume throughout that the density of y_i and the link function $h(\cdot)$ are correctly specified, but depart from a standard GLM setup by allowing for uncertainty in the specification of the linear predictor. Specifically, we partition the k regressors in two subsets, $X = [X_1 : X_2]$, where X_p is an $n \times k_p$ matrix with ith row equal to x_{ip}' ($p = 1, 2$) and $k_1 + k_2 = k$. The k_1 columns of X_1 contain the regressors which we want in the model on theoretical or other grounds (focus regressors in the terminology of Danilov and Magnus 2004), while the k_2 columns of X_2 contain the additional regressors of which we are less certain (auxiliary regressors). Stacking the linear predictors for the n observations on top of each other gives the n-vector $\eta(\beta) = X\beta = X_1\beta_1 + X_2\beta_2$, with $\beta = (\beta_1', \beta_2')'$, where β_1 is the vector of focus parameters and β_2 is the vector of auxiliary parameters.

In total, there are 2^{k_2} possible models that contain all focus regressors and arbitrary subsets of auxiliary regressors. We represent the jth model as a GLM of the form (1)–(2) with the added restriction $R_j'\beta_2 = 0$, where R_j denotes a $k_2 \times r_j$ matrix of rank $0 \le r_j \le k_2$ such that $R_j' = [I_{r_j} : 0]$ (or a column-permutation thereof) and I_{r_j} denotes the identity matrix of order r_j. The matrix R_j thus specifies which auxiliary regressors are excluded from the jth model and the scalar r_j denotes the number of excluded auxiliary variables. The fully restricted model that excludes all auxiliary regressors corresponds to the case when $R_j = I_{k_2}$ and $r_j = k_2$, while the unrestricted model that includes all auxiliary regressors corresponds to the case when $R_j = 0$ and $r_j = 0$.

As usual in the model averaging literature, we adopt an $\mathcal{M}$-closed framework where the unknown data-generation process (DGP) is included in the set of models considered by the investigator. Following the local misspecification framework (see, e.g., Hjort and Claeskens 2003a), we assume that the true value of the focus parameters β_1 is fixed while the true value of the auxiliary parameters β_2 is in a $\sqrt{n}$-shrinking neighborhood of zero. Although there is a debate about the realism of such assumption (see, e.g., Raftery and Zheng 2003, Ishwaran and Rao 2003, and Hjort and Claeskens 2003b), this framework has been commonly used to analyze the large sample behavior of a variety of estimators (see, e.g., Claeskens and Hjort 2003, Claeskens et al. 2006, Hansen 2014 and 2016, and Liu 2015). This framework has the great advantage of allowing the application of asymptotic model averaging theory as it ensures that all ML estimators are $\sqrt{n}$-consistent and have squared bias and variance both of order $O_p(n^{-1})$. On the other hand, in a standard asymptotic framework with a fixed value of β_2, we would always prefer the ML estimator of the unrestricted model because the ML estimator of the jth model may be inconsistent if the underlying constraint is not valid.

3 ML estimation

The classical approach to the estimation of GLMs is maximum likelihood. Given independent observations $\{(y_i, x_i')'\}_{i=1}^n$, the GLM loglikelihood is of the form

$$\ell(\beta) = c + \sum_{i=1}^n \left[\theta_i\, y_i - b(\theta_i)\right],$$

where c does not depend on β and the canonical parameter $\theta_i = \theta(\eta_i)$ depends on β through the linear predictor η_i. Since $x_i = (x_{i1}', x_{i2}')'$ and $\beta = (\beta_1', \beta_2')'$, the gradient of the loglikelihood (or likelihood score) is the k-vector $s(\beta)$ consisting of the following subvectors

$$s_p(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_p} = \sum_{i=1}^n v_i(\beta)\, [y_i - \mu_i(\beta)]\, x_{ip} \qquad (p = 1, 2),$$

where $v_i = d\theta/d\eta_i$. We also define a $k \times k$ matrix $H(\beta)$, which is equal to minus the Hessian of the loglikelihood and consists of the following submatrices

$$H_{pq}(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta_p \partial \beta_q'} = \sum_{i=1}^n \psi_i(\beta) x_{ip} x_{iq}' \qquad (p, q = 1, 2),$$

where $\psi_i = v_i^2\, \sigma_i^2 - \omega_i(y_i - \mu_i)$ and $\omega_i = d^2\theta/d\eta_i^2$, and a $k \times k$ matrix $I(\beta)$ (the Fisher information) consisting of the submatrices

$$I_{pq}(\beta) = \sum_{i=1}^n v_i(\beta)^2\, \sigma_i^2(\beta)\, x_{ip} x_{iq}' \qquad (p, q = 1, 2).$$

With a canonical link, these expressions simplify considerably as $\theta_i = \eta_i$, $v_i = 1$ and $\omega_i = 0$ for all observations, so $s_p(\beta) = \sum_{i=1}^n [y_i - \mu_i(\beta)]\, x_{ip}$ and $H_{pq}(\beta) = I_{pq}(\beta)$.

The ML estimator of β for the jth model maximizes the loglikelihood $\ell(\beta)$ subject to the constraint $R_j'\beta_2 = 0$ or, equivalently, solves the system of $k_1 + k_2 + r_j$ equations

$$\begin{aligned} 0 &= s_1(\beta), \\ 0 &= s_2(\beta) - R_j \nu_j, \\ 0 &= R_j'\beta_2, \end{aligned} \tag{3}$$

where ν_j denotes the r_j-vector of Lagrange multipliers associated with the constraint $R_j'\beta_2 = 0$. One issue in extending the WALS approach to the wider class of GLM is that, except when the elements of y are normally distributed, the system of likelihood equations (3) is nonlinear and has to be solved by some iterative scheme such as Newton-Raphson or the method of scoring. To address this issue we now introduce a class of one-step ML estimators that admit closed-form expressions and are asymptotically equivalent to the fully-iterated ML estimators.

3.1 One-step ML estimators

Given a starting value $\bar{\beta} = (\bar{\beta}_1', \bar{\beta}_2')'$, with properties to be discussed below, expanding the likelihood equations (3) around $\bar{\beta}$ yields the approximation

$$\begin{aligned}
0 &= \bar{s}_1 - \bar{H}_{11}(\beta_1 - \bar{\beta}_1) - \bar{H}_{12}(\beta_2 - \bar{\beta}_2), \\
0 &= \bar{s}_2 - \bar{H}_{21}(\beta_1 - \bar{\beta}_1) - \bar{H}_{22}(\beta_2 - \bar{\beta}_2) - R_j \nu_j, \\
0 &= R_j' \beta_2,
\end{aligned} \tag{4}$$

where $\bar{s}_p = s_p(\bar{\beta})$ and $\bar{H}_{pq} = H_{pq}(\bar{\beta})$, $p, q = 1, 2$. An estimator $\widetilde{\beta}_j$ that solves the linearized system of constrained likelihood equations (4) is called a one-step ML estimator of β under the jth model, as it corresponds to the first step of the Newton-Raphson method.

We first consider the unrestricted model where $R_j = 0$. Define the data transformations

$$\bar{y} = \bar{X}_1 \bar{\beta}_1 + \bar{X}_2 \bar{\beta}_2 + \bar{u}, \qquad \bar{X}_1 = \bar{\Psi}^{1/2} X_1, \qquad \bar{X}_2 = \bar{\Psi}^{1/2} X_2, \tag{5}$$

where $\bar{u} = \bar{\Psi}^{-1/2} \bar{V}(y - \bar{\mu})$, $\bar{\Psi} = \Psi(\bar{\beta})$ is an $n \times n$ diagonal matrix with ith diagonal element equal to $\psi_i(\bar{\beta})$, $\bar{V} = V(\bar{\beta})$ is an $n \times n$ diagonal matrix with ith diagonal element equal to $v_i(\bar{\beta})$, and $\bar{\mu} = \mu(\bar{\beta})$ is an n-vector with ith element equal to $\mu_i(\bar{\beta})$. Then, when $R_j = 0$, the solutions $\widetilde{\beta}_{1u}$ and $\widetilde{\beta}_{2u}$ to the linearized system of likelihood equations (4) can be written in closed form as

$$\widetilde{\beta}_{1u} = (\bar{X}_1' \bar{X}_1)^{-1} \bar{X}_1' \bar{y} - (\bar{X}_1' \bar{X}_1)^{-1} \bar{X}_1' \bar{X}_2 \widetilde{\beta}_{2u}, \qquad \widetilde{\beta}_{2u} = (\bar{X}_2' \bar{M}_1 \bar{X}_2)^{-1} \bar{X}_2' \bar{M}_1 \bar{y},$$

where $\bar{M}_1 = I_n - \bar{X}_1 (\bar{X}_1' \bar{X}_1)^{-1} \bar{X}_1'$ is a symmetric idempotent matrix of rank $n - k_1$. These expressions make it clear that the unrestricted one-step ML estimators $\widetilde{\beta}_{1u}$ and $\widetilde{\beta}_{2u}$ coincide numerically with the least squares coefficients in the linear regression of $\bar{y}$ on $\bar{X}_1$ and $\bar{X}_2$. Notice that, although the original regressors X_1 and X_2 are fixed (nonrandom), the transformed regressors $\bar{X}_1$ and $\bar{X}_2$ are in general random because they depend on $\bar{\beta}$ and y. In the canonical link case, the dependence on y disappears, as $\omega_i = 0$ for all i, but the dependence on $\bar{\beta}$ remains.

More generally, consider the one-step ML estimator for the jth model. After defining the symmetric and idempotent $k_2 \times k_2$ matrix

$$\bar{P}_j = \left(\frac{\bar{X}_2' \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2} R_j \left[R_j' \left(\frac{\bar{X}_2' \bar{M}_1 \bar{X}_2}{n} \right)^{-1} R_j \right]^{-1} R_j' \left(\frac{\bar{X}_2' \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2},$$

the $k_1 \times k_2$ matrix

$$\bar{Q} = \left(\frac{\bar{X}_1' \bar{X}_1}{n} \right)^{-1} \frac{\bar{X}_1' \bar{X}_2}{n} \left(\frac{\bar{X}_2' \bar{M}_1 \bar{X}_2}{n} \right)^{-1/2},$$

and the nonsingular transformation of the unrestricted one-step ML estimator $\widetilde{\beta}_{2u}$

$$\widetilde{\vartheta} = \left(\frac{\bar{X}_2' \bar{M}_1 \bar{X}_2}{n} \right)^{1/2} \widetilde{\beta}_{2u}, \tag{6}$$

we obtain the following generalization of Proposition 3.1 in Magnus and De Luca (2016).

Proposition 1 *The one-step ML estimators of β_1 and β_2 based on the jth model are*

$$\widetilde{\beta}_{1j} = \widetilde{\beta}_{1r} - \bar{Q}\bar{W}_j\widetilde{\vartheta}, \qquad \widetilde{\beta}_{2j} = \left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{-1/2}\bar{W}_j\widetilde{\vartheta},$$

where $\widetilde{\beta}_{1r} = (\bar{X}_1'\bar{X}_1)^{-1}\bar{X}_1'\bar{y}$ is the fully restricted one-step ML estimator of β_1 and $\bar{W}_j = I_{k_2} - \bar{P}_j$.

3.2 Asymptotic properties of one-step ML estimators

In what follows, to keep track of the sample size, we index all relevant data-dependent objects by n. Under the local misspecification framework, the auxiliary parameters are set equal to

$$\beta_{2n} = \frac{\delta}{\sqrt{n}}, \tag{7}$$

where δ is an unknown constant vector that represents the degree of model departure from the fully restricted model. Thus, the DGP depends on the sample size, with the sequence of true parameters $\beta_n = (\beta_1', \beta_{2n}')'$ converging to $\beta_* = (\beta_1', 0')'$ as $n \to \infty$.

The large-sample properties of the sequence $\{\widetilde{\beta}_{jn}\}$ of one-step ML estimators for the jth model depend crucially on the large-sample properties of the sequence $\{\bar{\beta}_n\}$ of starting values in the approximation (4). If $\bar{\beta}_n - \beta_n$ is $O_p(1/\sqrt{n})$, then $\widetilde{\beta}_{jn} - \beta_n$ is also $O_p(1/\sqrt{n})$ and has the same asymptotic distribution as the fully-iterated ML estimator of the jth model (see, e.g., Theorem 3.5 in Newey and McFadden 1994). In an $\mathcal{M}$-closed framework, where the DGP is included in the set of models considered by the investigator, a natural choice of starting value is the fully-iterated ML estimator based on the unrestricted model, as in this case $\bar{\beta}_n - \beta_n = O_p(1/\sqrt{n})$ under mild regularity conditions, irrespective of whether the local misspecification framework (7) is valid or not. These regularity conditions, spelled out in detail in Fahrmeir and Kaufmann (1985), essentially require the Fisher information $I_n(\cdot)$ to be continuous on an open neighborhood $\mathcal{B}$ of β_* and to diverge as the sample size grows. Under these conditions, $H_n(\cdot)/n$ and $I_n(\cdot)/n$ both converge in probability as $n \to \infty$, uniformly on $\mathcal{B}$, to a fixed (nonrandom), finite, symmetric and positive definite matrix $\mathcal{I}(\cdot)$.

The following result provides a convenient asymptotic approximation to the sampling distribution of one-step ML estimators under the local misspecification framework (7).

Proposition 2 *In addition to* (7), *assume that all regularity conditions in Fahrmeir and Kaufmann (1985) are satisfied. If $\bar{\beta}_n - \beta_n = O_p(1/\sqrt{n})$, then, as $n \to \infty$,*

$$\sqrt{n}(\widetilde{\beta}_{jn} - \beta_n) \Rightarrow \mathcal{N}\left(\begin{bmatrix} \mathcal{Q} \\ -\Omega_{22}^{1/2} \end{bmatrix} \mathcal{P}_j \Omega_{22}^{-1/2}\delta, \begin{bmatrix} \mathcal{I}_{11}^{-1} + \mathcal{Q}\mathcal{W}_j\mathcal{Q}' & -\mathcal{Q}\mathcal{W}_j\Omega_{22}^{1/2} \\ -\Omega_{22}^{1/2}\mathcal{W}_j\mathcal{Q}' & \Omega_{22}^{1/2}\mathcal{W}_j\Omega_{22}^{1/2} \end{bmatrix}\right),$$

where $\mathcal{I}_{pq}$ denotes the pqth submatrix of $\mathcal{I}(\beta_*)$, $\Omega_{22} = (\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12})^{-1}$, $\mathcal{Q} = \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}^{1/2}$, $\mathcal{P}_j = \Omega_{22}^{1/2}R_j(R_j'\Omega_{22}R_j)^{-1}R_j'\Omega_{22}^{1/2}$, and $\mathcal{W}_j = I_{k_2} - \mathcal{P}_j$.

The asymptotic distributions of the one-step ML estimators for the unrestricted and the fully restricted models are obtained as special cases by putting $R_j = 0$ and $R_j = I_{k_2}$, respectively. Proposition 2 is similar to Lemma 3.2 in Hjort and Claeskens (2003a) but differs because we consider the asymptotic distribution of the complete estimator $\widetilde{\beta}_{jn}$, including its r_j components restricted to be zero. Notice that

$$\sqrt{n}(\widetilde{\beta}_{jn} - \beta_*) = \sqrt{n}(\widetilde{\beta}_{jn} - \beta_n) + \begin{pmatrix} 0 \\ \delta \end{pmatrix},$$

so the two distributions only differ by a constant shift.

Three implications of Proposition 2 are worth noting. First, under the local misspecification framework, all estimators are consistent for β_*. If the jth model is correctly specified, that is, the constraint $R_j'\delta = 0$ is valid, then $\widetilde{\beta}_{jn}$ is asymptotically unbiased for β_n, since $\mathcal{P}_j\Omega_{22}^{-1/2}\delta = 0$, though not for β_*. However, if the constraint $R_j'\delta = 0$ is not valid, then $\widetilde{\beta}_{jn}$ is no longer asymptotically unbiased for β_n and its asymptotic bias may actually exceed that of estimators based on more parsimonious models.

Second, the asymptotic distribution of all estimators is normal and a comparison between the asymptotic variances of the restricted and unrestricted estimators yields

$$\mathrm{AV}(\widetilde{\beta}_{1un}) - \mathrm{AV}(\widetilde{\beta}_{1jn}) = \mathcal{Q}\mathcal{P}_j\mathcal{Q}'$$

and

$$\mathrm{AV}(\widetilde{\beta}_{2un}) - \mathrm{AV}(\widetilde{\beta}_{2jn}) = \Omega_{22}^{1/2}\mathcal{P}_j\Omega_{22}^{1/2},$$

which are two nonnegative definite matrices. Hence, irrespective of whether the constraint $R_j'\delta = 0$ is valid or not, the restricted estimators $\widetilde{\beta}_{1jn}$ and $\widetilde{\beta}_{2jn}$ are always asymptotically more precise (have smaller asymptotic variance) than the unrestricted estimators $\widetilde{\beta}_{1un}$ and $\widetilde{\beta}_{2un}$. This implies that the uncertainty about the choice of the auxiliary regressors gives rise to an asymptotic bias-precision trade-off in the estimation of β_n.

Third, it can be easily shown that $\sqrt{n}(\widetilde{\vartheta}_n - \vartheta_n) \Rightarrow \mathcal{N}(0, I_{k_2})$, where $\vartheta_n = \Omega_{22}^{-1/2}\beta_{2n}$. Further, $\widetilde{\vartheta}_n$ and $\widetilde{\beta}_{1rn}$ are asymptotically independent because their joint asymptotic distribution is normal with zero asymptotic covariance.

As we shall see in the next section, the results of Propositions 1 and 2 provide the key ingredients needed to extend the WALS approach to the wider class of GLMs.

4 WALS estimation

Our WALS approach to GLMs is a Bayesian combination of frequentist estimators that exploits a preliminary semiorthogonal transformation of the auxiliary regressors to reduce the computational burden required by exact model averaging estimation from the order 2^{k_2} to the order k_2. The parameters of each model are estimated by one-step ML based on a strictly frequentist approach, whereas the weighting scheme is based on a Bayesian approach to ensure desirable theoretical properties such as admissibility and a proper treatment of ignorance.

4.1 Scale and semiorthogonal transformations

To operationalize the WALS approach to GLM, we first transform the focus regressors in $\bar{X}_1 = \bar{\Psi}^{1/2}X_1$ by defining

$$\bar{Z}_1 = \bar{X}_1\bar{\Delta}_1, \qquad \bar{\gamma}_1 = \bar{\Delta}_1^{-1}\beta_1, \tag{8}$$

where $\bar{\Delta}_1$ is a diagonal $k_1 \times k_1$ matrix such that all diagonal elements of $\bar{Z}_1'\bar{Z}_1/n$ are equal to one. The only purpose of this transformation is to improve the numerical accuracy of inversion and eigenvalue routines. For the purposes of inference, this transformation is completely harmless because $\bar{Z}_1\bar{\gamma}_1 = \bar{X}_1\beta_1$, $I_n - \bar{Z}_1(\bar{Z}_1'\bar{Z}_1)^{-1}\bar{Z}_1' = \bar{M}_1$, and $\beta_1 = \bar{\Delta}_1\bar{\gamma}_1$.

Next we transform the auxiliary regressors in $\bar{X}_2 = \bar{\Psi}^{1/2}X_2$. Let $\bar{\Delta}_2$ be a diagonal $k_2 \times k_2$ matrix such that all diagonal elements of $\bar{\Xi} = \bar{\Delta}_2\bar{X}_2'\bar{M}_1\bar{X}_2\bar{\Delta}_2/n$ are equal to one. Notice that, unlike the matrix $\bar{\Delta}_1$, the matrix $\bar{\Delta}_2$ has the dual purpose of improving numerical accuracy and making the WALS estimator equivariant to scale transformations of the auxiliary regressors (De Luca and Magnus 2011). Since $\bar{\Xi}$ is a symmetric and positive definite matrix, we can apply the semiorthogonal transformation

$$\bar{Z}_2 = \bar{X}_2\bar{\Delta}_2\bar{\Xi}^{-1/2}, \qquad \bar{\gamma}_{2n} = \bar{\Xi}^{1/2}\bar{\Delta}_2^{-1}\beta_{2n}, \tag{9}$$

which implies that $\bar{Z}_2'\bar{M}_1\bar{Z}_2/n = I_{k_2}$, $\bar{Z}_2\bar{\gamma}_{2n} = \bar{X}_2\beta_{2n}$, and $\beta_{2n} = \bar{\Delta}_2\bar{\Xi}^{-1/2}\bar{\gamma}_{2n}$.

The transformations (8) and (9) present two important differences with respect to those employed in the WALS approach to linear models. The first difference is that, with a view toward asymptotic analysis, we have normalized all relevant matrices by n to ensure that they remain stable when the sample size becomes arbitrarily large.

The second difference lies in the semiorthogonal transformation (9) where we now avoid possible discontinuities in the eigenvectors and eigenprojections of the matrix $\bar{\Xi}$ by exploiting the continuity of the eigenvalues and the total eigenprojections. As shown in Appendix B, this ensures that $\bar{\Xi}^{1/2}$, $\bar{\Xi}^{-1}$, and $\bar{\Xi}^{-1/2}$ are continuous matrix functions, as long as $\bar{\Xi}$ is continuous and positive definite. The large-sample probability limits of the random objects in (8) and (9) then follow easily. Since $\operatorname{plim} \bar{X}_1'\bar{X}_1/n = \operatorname{plim} \bar{H}_{11}/n = \mathcal{I}_{11}$, the matrix $\bar{\Delta}_1$ converges in probability as $n \to \infty$ to a diagonal nonrandom matrix Δ_1 with diagonal elements equal to the inverse of the square root of the diagonal elements of $\mathcal{I}_{11}$, so that

$$\operatorname{plim} \frac{\bar{Z}_1'\bar{Z}_1}{n} = \Delta_1\mathcal{I}_{11}\Delta_1 = \mathcal{J}_{11}.$$

Similarly, because of continuity of the inverse of a nonsingular matrix, the scaling matrix $\bar{\Delta}_2$ converges in probability to a diagonal nonrandom matrix Δ_2 with diagonal elements equal to the inverse of the square root of the diagonal elements of Ω_{22}^{-1}, so that

$$\operatorname{plim} \bar{\Xi} = \Delta_2\Omega_{22}^{-1}\Delta_2 = \Xi.$$

Moreover, the continuity of $\bar{\Xi}^{-1/2}$ now implies that

$$\operatorname{plim} \frac{\bar{Z}_1'\bar{Z}_2}{n} = \Delta_1\mathcal{I}_{12}\Delta_2\Xi^{-1/2} = \mathcal{J}_{12}$$

and

$$\operatorname{plim} \frac{\bar{Z}_2'\bar{Z}_2}{n} = \Xi^{-1/2}\Delta_2\mathcal{I}_{22}\Delta_2\Xi^{-1/2} = \mathcal{J}_{22},$$

so that $\mathcal{J}_{22} - \mathcal{J}_{21}\mathcal{J}_{11}^{-1}\mathcal{J}_{12} = I_{k_2}$.

4.2 One-step ML estimation of the transformed models

Since $\bar{Z}_1\bar{\gamma}_1 = \bar{X}_1\beta_1$ and $\bar{Z}_2\bar{\gamma}_{2n} = \bar{X}_2\beta_{2n}$, we can rewrite the unrestricted model as a GLM of the form (1)–(2) with linear predictor $\eta = \bar{Z}_1\bar{\gamma}_1 + \bar{Z}_2\bar{\gamma}_{2n}$. This equivalent representation is convenient because it implies that $\bar{Z}_2'\bar{M}_1\bar{Z}_2/n = I_{k_2}$. It then follows from Proposition 1 that the one-step ML estimators for the jth model are given by

$$\widetilde{\gamma}_{1jn} = \widetilde{\gamma}_{1rn} - \bar{D}W_j\widetilde{\gamma}_{2un}, \qquad \widetilde{\gamma}_{2jn} = W_j\widetilde{\gamma}_{2un}, \tag{10}$$

where $\widetilde{\gamma}_{1rn} = (\bar{Z}_1'\bar{Z}_1)^{-1}\bar{Z}_1'\bar{y}$, $\widetilde{\gamma}_{2un} = \bar{Z}_2'\bar{M}_1\bar{y}/n$, $\bar{D} = (\bar{Z}_1'\bar{Z}_1)^{-1}\bar{Z}_1'\bar{Z}_2$, $W_j = I_{k_2} - P_j$, and $P_j = R_jR_j'$.

Further, letting $\gamma_n = (\gamma_1{}', \gamma_{2n}{}')'$ with $\gamma_1 = \Delta_1^{-1}\beta_1$ and $\gamma_{2n} = \Xi^{1/2}\Delta_2^{-1}\beta_{2n}$, Proposition 2 also implies

$$\sqrt{n}(\widetilde{\gamma}_{jn} - \gamma_n) \Rightarrow \mathcal{N}\left(\begin{bmatrix} \mathcal{D} \\ -I_{k_2} \end{bmatrix} P_j d, \begin{bmatrix} \mathcal{J}_{11}^{-1} + \mathcal{D}W_j\mathcal{D}' & -\mathcal{D}W_j \\ -W_j\mathcal{D}' & W_j \end{bmatrix}\right), \tag{11}$$

where $d = \Xi^{1/2}\Delta_2^{-1}\delta$ and $\mathcal{D} = \operatorname{plim}\bar{D} = \mathcal{J}_{11}^{-1}\mathcal{J}_{12}$. Thus, as a direct consequence of (9), the matrix W_j now reduces to a nonrandom diagonal matrix with $k_2 - r_j$ ones and r_j zeros on its main diagonal. More precisely, the hth diagonal element of W_j is equal to zero if the hth component of γ_{2n} is constrained to be zero, and is equal to one otherwise. All models that include the hth column of $\bar{Z}_2$ as a regressor will therefore have the same estimator of the hth component of γ_{2n}, namely the hth component of $\widetilde{\gamma}_{2un}$. The components of $\widetilde{\gamma}_{2un}$ are asymptotically independent as their joint asymptotic distribution is normal with zero asymptotic covariance.

4.3 Equivalence theorem

We next consider the model averaging estimators of γ_1 and γ_{2n}

$$\widehat{\gamma}_{1n} = \sum_{j=1}^{2^{k_2}} \lambda_j\widetilde{\gamma}_{1jn}, \qquad \widehat{\gamma}_{2n} = \sum_{j=1}^{2^{k_2}} \lambda_j\widetilde{\gamma}_{2jn},$$

where the λ_j are data-dependent model weights satisfying the restrictions

$$0 \le \lambda_j \le 1, \qquad \sum_{j=1}^{2^{k_2}} \lambda_j = 1, \qquad \lambda_j = \lambda_j(\sqrt{n}\widetilde{\gamma}_{2un}). \tag{12}$$

From (10) we get

$$\widehat{\gamma}_{1n} = \widetilde{\gamma}_{1rn} - \bar{D}W\widetilde{\gamma}_{2un}, \qquad \widehat{\gamma}_{2n} = W\widetilde{\gamma}_{2un}, \tag{13}$$

where $W = \sum_{j=1}^{2^{k_2}} \lambda_jW_j$ is a $k_2 \times k_2$ random diagonal matrix (because the λ_j are random) and the random vector $W\widetilde{\gamma}_{2un}$ is asymptotically independent of $\widetilde{\gamma}_{1rn}$.

The following proposition extends the finite-sample results of Magnus and Durbin (1999) and Danilov and Magnus (2004) and the large-sample results of Zou et al. (2007), which only cover linear models, and motivates the WALS approach to GLMs.

Proposition 3 (Asymptotic Equivalence Theorem for GLMs) *Under the regularity conditions* (12)*, the asymptotic bias* (AB) *and the asymptotic variance* (AV) *of the WALS estimator*

$\widehat{\gamma}_{1n}$ *of* γ_1 *are respectively related to the asymptotic bias and the asymptotic variance of the WALS estimator* $\widehat{\gamma}_{2n}$ *of* γ_{2n} *by the relationships*

$$\mathrm{AB}(\widehat{\gamma}_{1n}) = -\mathcal{D}\,\mathrm{AB}(\widehat{\gamma}_{2n}), \qquad \mathrm{AV}(\widehat{\gamma}_{1n}) = \mathcal{J}_{11}^{-1} + \mathcal{D}\,\mathrm{AV}(\widehat{\gamma}_{2n})\mathcal{D}'.$$

Hence, the asymptotic mean squared errors (AMSE) *of* $\widehat{\gamma}_{1n}$ *and* $\widehat{\gamma}_{2n}$ *are linked by the relationship*

$$\mathrm{AMSE}(\widehat{\gamma}_{1n}) = \mathcal{J}_{11}^{-1} + \mathcal{D}\,\mathrm{AMSE}(\widehat{\gamma}_{2n})\mathcal{D}'.$$

The equivalence theorem implies that the AMSE of the WALS estimator $\widehat{\gamma}_{1n}$ depends on the AMSE of the less complicated estimator $\widehat{\gamma}_{2n}$. This means that, if we can choose the model weights λ_j such that $\widehat{\gamma}_{2n}$ is a 'good' estimator of γ_{2n}, then the same λ_j will also provide a 'good' estimator of γ_1. The problem of choosing the model weights optimally is much simplified by the fact that W is a diagonal matrix whose diagonal elements w_h are linear combinations of the λ_j. The computational burden of our model averaging estimator is therefore of order k_2, as we only need to determine the set of k_2 WALS weights w_h, not the considerably larger set of 2^{k_2} model weights λ_j.

4.4 Bayesian weighting scheme and choice of priors

Since the WALS weights w_h lie between zero and one, the components of $\widehat{\gamma}_{2n}$ are shrinkage estimators of the components of γ_{2n}. We also know that the components of $\widetilde{\gamma}_{2un}$ are asymptotically independent, each with an asymptotically normal distribution. Thus, if we strengthen the third regularity condition in (12) and assume that each w_h depends only on the hth component of $\sqrt{n}\widehat{\gamma}_{2un}$, then the shrinkage estimators in $\widehat{\gamma}_{2n}$ will also be asymptotically independent. This additional assumption is convenient because our k_2-dimensional problem then reduces to k_2 (identical) one-dimensional problems of the following type: given a shrinkage estimator $m(x) = w(x)x$ of a scalar parameter γ, we want to determine the scalar weight $w(x)$ such that the estimator $m(x)$ has minimum MSE by only using the information that $x \sim \mathcal{N}(\gamma, 1)$. This is the normal location problem studied and refined in a finite-sample context by Magnus (2002), Kumar and Magnus (2013), and Magnus and De Luca (2016), and now extended to the asymptotic distribution of $\widehat{\gamma}_{2n}$.

Our search for an optimal weighting scheme can be developed along frequentist or Bayesian lines. In WALS we prefer a Bayesian weighting scheme because it leads to an admissible shrinkage estimator of γ. The issue of how to choose the prior for this Bayesian step has recently been addressed by Magnus and De Luca (2016) who focused on the family of reflected generalized gamma distributions that satisfy a number of conditions for robustness and proper treatment of ignorance. These priors have densities of the form

$$\pi(\gamma) = \frac{qc}{2}|\gamma|^{-(1-q)}e^{-c|\gamma|^q}, \tag{14}$$

with $c = 0.9377$ and $q = 0.7995$ corresponding to the optimal Subbotin prior, and $c = \log 2$ and $q = 0.8876$ corresponding to the optimal reflected Weibull prior. The Subbotin prior is preferred in terms of robustness, while the reflected Weibull prior is preferred in terms of minimax regret (Magnus and De Luca 2016). In both cases, the moments of the resulting posterior distribution need to be approximated by numeric integration techniques. Closed-form expressions for the posterior mean and the posterior variance are available only under the Laplace prior, corresponding to $c = \log 2$ and $q = 1$ (see Theorem 1 in Magnus et al. 2010), but this choice is neither robust nor optimal in terms of minimax regret.

4.5 One-step and iterative WALS estimates

Letting m be the k_2-vector of posterior means and Σ the $k_2 \times k_2$ diagonal matrix with the posterior variances as diagonal elements, we can now define the one-step WALS estimators of γ_1 and γ_{2n} as

$$\widehat{\gamma}_{1n} = \widetilde{\gamma}_{1rn} - \bar{D}m, \qquad \widehat{\gamma}_{2n} = m.$$

Consistent estimators of their asymptotic variances are

$$\widehat{\mathrm{AV}}(\widehat{\gamma}_{1n}) = \left(\frac{\bar{Z}_1'\bar{Z}_1}{n}\right)^{-1} + \bar{D}\Sigma\bar{D}', \qquad \widehat{\mathrm{AV}}(\widehat{\gamma}_{2n}) = \Sigma.$$

The one-step WALS estimator of the original parameters β_1 and β_{2n} are then given by

$$\widehat{\beta}_{1n} = \bar{\Delta}_1\widehat{\gamma}_{1n}, \qquad \widehat{\beta}_{2n} = \bar{\Delta}_2\bar{\Xi}^{-1/2}\widehat{\gamma}_{2n}, \tag{15}$$

and their asymptotic variances can be estimated consistently by

$$\widehat{\mathrm{AV}}(\widehat{\beta}_{1n}) = \bar{\Delta}_1\widehat{\mathrm{AV}}(\widehat{\gamma}_{1n})\bar{\Delta}_1', \qquad \widehat{\mathrm{AV}}(\widehat{\beta}_{2n}) = \bar{\Delta}_2\bar{\Xi}^{-1/2}\widehat{\mathrm{AV}}(\widehat{\gamma}_{2n})\bar{\Xi}^{-1/2}\bar{\Delta}_2'. \tag{16}$$

One possible drawback of the one-step WALS procedure could be its dependence on the starting value $\bar{\beta}$. To address this issue we also consider an iterative procedure that repeatedly updates the starting value $\bar{\beta}$ using the one-step WALS estimates from the previous iteration until some convergence criterion is satisfied. The rationale behind this iterative procedure is that, as the number of iterations increases, the sequence of recursive applications of the one-step estimator of the jth model converges to the corresponding fully-iterated ML estimator (Robinson 1988, Theorem 2). Thus, when $\bar{\beta}$ is a $\sqrt{n}$-consistent estimator of β_*, there are reasons to believe that the iterative WALS estimator provides a good approximation to a weighted average over all possible models of the fully-iterated ML estimators. In what follows we shall assess the dependence of the one-step and iterative WALS estimates on alternative choices of the starting value $\bar{\beta}$, both by an empirical study and by a Monte Carlo experiment.

4.6 Estimating smooth functions of the model parameters

In the context of GLMs, inference is usually sought for a smooth, but possibly nonlinear, real-valued function $g(\beta; x)$ of the model parameters β at some value x of the regressors. Examples include the probability of success in a binary logit model or the marginal effect of a given regressor. In this section, we thus focus on the problem of estimating $g(\beta; x)$ when there is uncertainty about the choice of the auxiliary regressors.

From a frequentist perspective, ML estimation of each possible model yields a set of 2^{k_2} conditional ML estimates $\widehat{\beta}_j$, from which we obtain the conditional ML estimates $\widehat{g}_j = g(\widehat{\beta}_j; x)$ of $g(\beta; x)$. The key issue is how to best combine them to construct an unconditional estimate of $g(\beta; x)$ that incorporates the uncertainty due to both the model selection and the model estimation steps. The standard FMA solution is an estimator of the form

$$\widehat{g}_{ma} = \sum_{j=1}^{2^{k_2}} \lambda_j^* \widehat{g}_j, \tag{17}$$

where the λ_j^* are model weights chosen on the basis of some optimality criterion (see, e.g., Hjort and Claeskens 2003a). BMA estimators have a similar form, that is, they are a weighted average of the means of the conditional posterior distributions of $g(\beta; x)$ under each possible model with weights equal to the posterior model probabilities (see, e.g., Hoeting et al. 1999).

Unfortunately, in WALS we cannot construct the model averaging estimator in (17) due to lack of information on the $\widehat{g}_j$ and the λ_j^*. This is a consequence of the semiorthogonal transformation (9) which leads to important simplifications when estimating β, but also implies some loss of flexibility compared to standard FMA and BMA approaches. Here loss of flexibility means that we can only compute a model averaging estimator $\widehat{\beta}$ of β, that is

$$\widehat{\beta} = \sum_{j=1}^{2^{k_2}} \lambda_j \widehat{\beta}_j, \tag{18}$$

on the basis of which we then obtain a plug-in estimator $\widehat{g}_{pi} = g(\widehat{\beta}; x)$ of $g(\beta; x)$. Thus, instead of averaging over nonlinear transformations of the ML estimators, we can only apply a nonlinear transformation of the model averaging estimator of β.

These two classes of estimators are likely to differ as a consequence of both Jensen's inequality and possible differences in model weights. Apart from Koenker (2005, Section 5.5), little is known about the statistical properties of one class relative to the other. Koenker discusses not precisely our question, but the related issue of comparing weighted averages of argmins and argmins of weighted averages in the context of quantile regressions. A key result from his analysis is that these two classes of estimators reach the same efficiency bound, but that the associated sets of optimal weights are in general different. This result suggests that when the model weights are determined on the basis of a well-defined criterion neither of the two estimators is expected to dominate the other. To shed some light on this topic, our empirical illustration in Section 5 focuses on estimating the probability of success in a binary logit model, which allows us to compare the performance of the plug-in estimator obtained in the WALS approach with the model averaging estimators obtained in standard BMA and FMA approaches.

5 Empirical illustration

We illustrate the WALS approach to GLMs by studying attrition in the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national household panel survey which covers about 85,000 individuals aged 50+, and their possibly younger partners, in nineteen countries of Continental Europe and Israel.

5.1 Data and model specification

Our data are from release 5.0 of SHARE. For detailed information on sampling design, eligibility rules, sample composition, country coverage, and fieldwork procedures, we refer to Malter and Börsch-Supan (2015). Here we only discuss a few issues that are important for the selection of the sample used in our empirical illustration. First, although five waves of SHARE are currently available, we focus on attrition between the first two waves (2004–05 and 2006–07) to avoid modeling differences in participation probabilities between the baseline sample drawn in the first wave and the refreshment samples drawn in subsequent waves. Second, since participation decisions of individuals belonging to the same household are likely to be correlated, we confine attention to one person per household, the so-called 'household respondent'. Third, to reduce issues of sample representativeness for certain population groups, we further restrict our sample to household respondents between 50 and 85 years old in 2004 and living in private households.

After dropping another 6% of the sample because of item nonresponse on the regressors of interest, our working sample contains 17,051 individuals, with national samples ranging from a minimum of 620 individuals for Switzerland to a maximum of 2,323 individuals for Belgium. The participation rate between the first two waves of SHARE ranges from a minimum of 55% in Germany to a maximum of 86% in Greece, and is 71% on average. For the purpose of this empirical illustration we focus on France (1,822 individuals with a participation rate of 68%), where the problem of uncertainty concerning the choice of regressors appears to be particularly relevant. Corresponding analyses for the other countries are available from the authors upon request.

Our outcome of interest is a binary indicator y_i, which equals 1 if a household participating in the baseline survey also agrees to participate in the second wave of SHARE, and equals 0 otherwise. We model the observed data $y_1, \dots, y_n$ as independent binary random variables, each having a Bernoulli distribution with probability of success $\pi_i = \Pr(y_i = 1) = [1 + \exp(-\eta_i)]^{-1}$, where $\eta_i = x_i'\beta$. The set of focus regressors in x_i includes a constant term, a second-order polynomial in age, a binary indicator for being a female fully interacted with the polynomial in age, and four binary indicators for other socio-economic characteristics of the household respondent (living with a spouse/partner, living in a big city, having at least a high school degree, and being employed), while the set of auxiliary regressors includes measures of physical and mental health, cognitive functioning, and social activities of the respondent, plus demographic characteristics of the partner and of the interviewer. In total we select eight auxiliary regressors, which results in $2^8 = 256$ possible models. Table 1 shows definitions and summary statistics for all the variables considered.

5.2 Estimation methods

Our empirical illustration has three purposes. First, we want to compare our approach with other popular strictly Bayesian (BMA) and strictly frequentist (FMA) model averaging procedures. Second, we want to investigate the robustness of the various model averaging approaches to key features of the underlying weighting scheme, including the choice of prior distributions for the weights used in WALS and BMA, and the choice of optimality criteria for the weights used in FMA. Third, we want to assess the sensitivity of one-step and iterative WALS estimates to the choice of the starting value. In the remaining of this section, we briefly describe the three model averaging approaches implemented in our empirical study. Stata routines for WALS, BMA and FMA estimation are available from the authors upon request.

As starting value for WALS we consider the restricted and the unrestricted ML estimates. After implementing the preliminary data transformations in (5), with $\mu_i = \pi_i$, $\sigma_i^2 = \pi_i(1 - \pi_i)$, $v_i = 1$, and $\omega_i = 0$, the one-step estimates are computed through the standard WALS procedure for linear models by setting the error variance equal to one. As priors on the transformed parameter γ, we consider the Subbotin, Weibull and Laplace priors discussed in Section 4.4. For the Subbotin and Weibull priors, we approximate the indefinite integrals needed for the first two moments of the posterior distribution using Gauss-Laguerre quadrature methods with $1,000$ points. To compute the iterative WALS estimates, we repeatedly update the starting value using the estimates from the previous iteration until the relative differences in the vectors of coefficients and their standard errors are both lower than the tolerance value of 10^{-6}.

For the BMA approach we compute a weighted average of the conditional estimates for each possible model with weights equal to the posterior model probabilities. Contrary to WALS, which uses priors only on the transformed parameters γ, BMA requires two types of priors: one on the model space and one on the parameters of each model (see, e.g., Hoeting et al. 1999). Our BMA implementation uses a uniform prior on the model space and conjugate priors for the parameters of each model. The first choice implies that all models are equally likely a priori, so their posterior model probabilities depend only on the marginal likelihood for the various models, not on the prior weight assigned to each of them. Following Chen and Ibrahim (2003), our conjugate prior for the free parameters β_j of the jth model is proportional to $\exp\left[\bar{a}(\bar{y}'\theta(\beta_j) - \iota_n' b(\theta(\beta_j)))\right]$, where $\bar{y}$ is an n-vector of prior parameters that specifies the prior predictions for the marginal means of the outcome, the positive scalar $\bar{a}$ is a prior parameter that quantifies the strength of our prior belief in $\bar{y}$, $\theta(\beta_j) = (\theta_1(\beta_j), \ldots, \theta_n(\beta_j))$ is the n-vector of canonical parameters in the jth model, and ι_n is an n-vector of ones. As shown by Chen et al. (2008), this family of priors is attractive because the posterior model probabilities can be estimated by a computationally convenient Markov Chain Monte Carlo (MCMC) method that requires drawing only two MCMC samples: one from the posterior distribution and one from the prior distribution of the parameters under the unrestricted model. In our application, we employ two MCMC samples of $10,000$ draws, after a 'burn-in sample' of $5,000$ draws. To ensure that all parameters have a zero prior mode we set all elements of $\bar{y}$ equal to 0.5. We also asses how BMA estimates change as the prior becomes less informative by considering three different values of $\bar{a}$, namely 0.10, 0.05, and 0.01.

For the FMA approach we compute weighted averages of the conditional ML estimates for each possible model with weights equal to, respectively, the smoothed Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the focused information criterion (FIC). The use of FMA estimators with smoothed AIC and BIC weights was originally proposed by Buckland et al. (1997) and is common in the context of BMA estimation (see, e.g., Raftery 1996 and Clyde 2000). Although debate over the choice of an optimal information criterion is still open, AIC and BIC are known to be two extreme strategies favoring, respectively, more and less complicated model structures. The smoothed FIC weights proposed by Hjort and Claeskens (2003a) are a little different, as they also depend on the specific parameter $g(\beta; x)$ to be estimated. Since the FIC score for the jth model is an unbiased estimator of the AMSE of the underlying ML estimator of $g(\beta; x)$, the smoothed FIC weighting scheme assigns relatively higher weights to models with relatively lower FIC scores. In our empirical illustration, we compute FMA estimates with smoothed FIC weights related to the participation probabilities of representative males and females aged between 50 and 85 years.

5.3 Estimation results

Table 2 presents the estimates of our logit models for the probability of survey participation in the second wave of the French SHARE, conditional on participation in the first wave. The table compares estimates and standard errors of the focus parameters for ten estimators: the restricted and unrestricted ML estimators, two FMA estimators, three BMA estimators, two one-step WALS estimators, and the iterative WALS estimator. For brevity, we only report the FMA estimates based on the smoothed AIC and BIC weights, and the WALS estimates based on the Weibull prior.

Except for the coefficient on the dummy variable for living with a partner (`Couple`), our results show no differences in the signs of the estimated associations across estimation methods. However, the size of the coefficients and the standard errors reveal nonnegligible differences. The importance of model uncertainty in the present application is confirmed by the fact that model weights from the FMA and the BMA approaches are clearly spread out across several models. The best-performing model changes depending on the weighting scheme, but the largest model weight is always lower than 0.18 for FMA and 0.14 for BMA. In WALS, this type of information is not available because we estimate only $k_2 = 8$ linear combinations of the $2^{k_2} = 256$ model weights.

The FMA, BMA and WALS estimates are often in-between the restricted and the unrestricted ML estimates, but generally closer to the latter. As for WALS, we find that the one-step estimates are rather insensitive to the choice of the starting value. The one-step WALS with starting value $\bar{\beta} = \widehat{\beta}_r$ always has smaller standard errors than the one-step WALS with starting value $\bar{\beta} = \widehat{\beta}_u$, but they do not differ much from the FMA and BMA standard errors and are always lower than unrestricted ML standard errors. In the iterative version of WALS, different starting values affect only the number of iterations needed for convergence (4 with $\bar{\beta} = \widehat{\beta}_u$ and 5 with $\bar{\beta} = \widehat{\beta}_r$), but not the estimated coefficients and standard errors.

Figures 1–3 plot the gender-specific age-profiles of participation probabilities estimated from the ML, FMA and BMA approaches, along with the estimates and the one-standard error bands from the iterative WALS approach. For the FMA approach in Figure 2, we also illustrate the estimates obtained with the smoothed FIC weights. Each point of the estimated age-profiles corresponds to the participation probabilities of a representative male and a representative female aged a years. For ML and WALS we compute plug-in estimates, whereas the BMA and FMA estimates are computed according to (17). The WALS standard errors are computed by the delta method. The restricted and unrestricted ML estimates differ considerably, whereas the WALS, BMA and FMA estimates are remarkably similar and close to the unrestricted ML estimates. Particularly striking is the similarity of the estimates from iterative WALS, FMA with smoothed FIC weights and BMA with prior parameter $\bar{a} = 0.05$, suggesting that the results from WALS are comparable to those from other popular model averaging methods. Our approach is also robust to different choices of the starting value and to different choices of prior on the transformed parameters. An important advantage of WALS compared to other approaches is that it can be obtained in negligible computing time.

6 Monte Carlo simulations

This section present the results of a Monte Carlo experiment which compares the finite-sample performance of the various ML and model averaging estimators within a realistic simulation setup based on the empirical study of survey participation described in the previous section.

More precisely, we set the parameters of the DGP equal to the unrestricted ML estimates $\widehat{\beta}_u$ presented in Table 2 and consider four simulation designs corresponding to alternative sample sizes (n = 100, 400, 900 and 1,600). In the tth design ($t = 1, \dots, 4$), we use simple random sampling with replacement to draw subsamples $X_t = [X_{1t} : X_{2t}]$ of size n_t from the original design matrix X with 1,822 observations. We then simulate the outcome y_{it} for the ith observation of the tth subsample by a pseudo-random draw from a Bernoulli distribution with probability of success $\pi_{it} = [1 + \exp(-x'_{it}\beta_t)]^{-1}$, where $\beta_t = (\widehat{\beta}'_{1u}, \sqrt{n/n_t}\,\widehat{\beta}'_{2u})'$.

We focus on estimating the survey participation probabilities π_m and π_f of a representative male and a representative female with 70 years of age. Under our Monte Carlo design, $\pi_m = 0.7301$

and $\pi_f = 0.7522$. Summaries of the sampling distribution of each estimator are approximated using 1,000 Monte Carlo replications.

Table 3 presents the bias, standard error (SE) and root mean squared error (RMSE) of the various ML and model averaging estimators. For WALS we only report the estimates based on the Weibull prior because Subbotin and Laplace priors yield very similar results.

Our results show a clear bias-precision trade-off in the choice between the two ML estimators. Since the unrestricted model is always correctly specified, the bias of its ML estimator is close to zero for any n. In small samples ($n = 100$), the restricted ML estimator is considerably biased. However, as n increases, its bias converges to zero because the auxiliary parameters of the DGP satisfy the local misspecification framework. A comparison of the SE suggests that the restricted ML estimator is always more precise than the unrestricted, but the reduction in the variance does not always compensate for the associated bias. Thus, in most simulation designs, the unrestricted ML estimator has lower RMSE than the restricted ML estimator.

BMA, FMA and WALS estimators always dominate the restricted and the unrestricted ML estimators in terms of RMSE. In all simulation designs, the FMA-FIC estimator has considerably lower RMSE than the FMA-AIC and FMA-BIC estimators mainly because of its higher precision. In contrast, the RMSE of BMA and WALS estimators depends on the sample size. For BMA, the preferred prior parameter is $\bar{a} = 0.10$ when $n \leq 400$ and $\bar{a} = 0.05$ when $n > 400$. For WALS, the iterative estimator performs slightly better than the one-step estimators when $n > 100$. In small samples, the one-step estimator with starting value $\bar{\beta} = \widehat{\beta}_r$ is the most precise. The three types of model averaging estimator always have similar finite-sample performance. RMSE comparisons favor BMA over WALS and WALS over FMA-FIC when $n = 100$, and FMA-FIC over WALS and WALS over BMA when $n > 100$, but the differences are always small.

7 Conclusions

This paper extends the WALS approach for dealing with uncertainty about the specification of the linear predictor from the linear Gaussian regression model to the wider class of GLMs. Our one-step WALS estimator for GLMs consists of three logical stages. Based on a strictly frequentist approach, we first estimate the parameters of each GLM by one-step ML which is numerically equivalent to least squares in a regression on transformed data for the outcome and the regressors. Second, we use a semiorthogonal transformation which reduces the computational burden required by model averaging estimation from the order 2^{k_2} to the order k_2. Third, we estimate the required k_2 linear combinations of the 2^{k_2} model weights by a Bayesian approach which allows a proper treatment of ignorance in the choice of the prior, satisfies other theoretical properties such as admissibility and robustness, and is optimal in terms of minimax regret. Since the one-step ML estimator depends on an arbitrarily chosen starting value, we also consider an iterative WALS estimator which repeatedly updates the starting value with the one-step WALS estimates from the previous iteration until convergence.

Results from both an empirical illustration and a related Model Carlo experiment on attrition in the first two waves of the French SHARE show that the one-step and iterative WALS estimators outperform standard ML estimators of the restricted and unrestricted models. The finite-sample performance of our estimators are remarkably similar to those of the FMA estimator with smoothed FIC weights (Hjort and Claeskens 2003a) and the BMA estimator with conjugate priors for GLMs (Chen and Ibrahim 2003; and Chen et al. 2008). The key advantage with respect to these alternative model averaging procedures is that WALS estimates can be computed in negligible computing time. This computational advantage is likely to be important in empirical applications where estimation of all possible models is not feasible. In addition, WALS is robust to different choices of the starting values and different choices of the priors for the Bayesian weighting scheme.

In this paper we focused on WALS estimation of GLMs for a scalar outcome of interest, but our model averaging procedure could be further extended in several important directions. First, an extension of the WALS approach to GMLs for multivariate outcomes would open the way to a larger variety of models, such as seemingly unrelated regression equations, and ordered, multinomial, and conditional logit and probit models. Second, under regularity conditions analogous to those required for ML estimation, the asymptotic WALS theory developed here could also be extended to general M-estimators of linear index models. Third, in addition to standard regularity conditions for GLMs, our asymptotic WALS theory is based on an $\mathcal{M}$-closed local misspecification framework, where the unknown DGP is included in the set of models considered by the investigator and the biases of the underlying ML estimators gradually shrink to zero with sample size at the rate $n^{-1/2}$. These assumptions ensure that all ML estimators are $\sqrt{n}$-consistent and that there exists a well-defined bias-precision trade-off in their asymptotic distributions. Despite the significant progresses made in the recent years, we believe that considerable theoretical work is still required to extend the existing model-averaging techniques to more general frameworks. This is a challenging and important line for future research.

Table 1: Definitions and summary statistics for the variables in France

Variable	Description	Mean	SD	Min	Max
Part	Dummy participation in w2	0.68	0.47	0	1
Age	Age of HR in 2004	64.37	9.99	50	85
Age^2/10	Squared age of HR divided by 10	4243.07	1320.22	2500	7225
Female	Dummy female HR	0.53	0.50	0	1
Female × Age	INT Female - Age	34.76	33.44	0	85
Female × Age^2/10	INT Female - Age^2/10	2325.39	2397.71	0	7225
Couple	Dummy living with a partner	0.59	0.49	0	1
Big City	Dummy living in a big city	0.43	0.50	0	1
High Education	Dummy high education	0.57	0.50	0	1
Employed	Dummy being employed	0.28	0.45	0	1
Good SRH	Dummy for good SRH	0.68	0.47	0	1
Doctor	Number of visits to medical doctor	6.85	7.19	0	98
Euro-D	Euro-D depression index	2.80	2.31	0	12
Recall	Score of recall tests	7.47	3.29	0	18
Social Activities	Number of social activities	0.80	1.00	0	6
Couple × Age Partner	INT Couple - Age of HR's partner	36.23	31.30	0	90
IV Female	Dummy female interviewer	0.76	0.43	0	1
IV Age	Age of interviewer in 2004	51.03	7.54	19	80

Notes: Sample size is 1,822 individuals. 'Part' is our binary outcome variable. Focus and auxiliary regressors are listed, respectively, in the second and the third panels. HR means 'household respondent', INT means 'interaction term', SRH means 'self-reported health', and IV means 'interviewer'. In estimation we center 'Age', 'Age of Partner', and 'IV Age' at 50, 'Doctor' at 5, 'Euro-D' at 3, 'Recall' at 9, and 'Social Activities' at 1.

Table 2: Estimates and standard errors of the focus parameters in the logit model for the probability of participation in the second wave of the French SHARE panel conditional on participation in the first wave

	ML		FMA		BMA			WALS		
Regressors	$\widehat{\beta}_r$	$\widehat{\beta}_u$	BIC	AIC	$\bar{a} = 0.01$	$\bar{a} = 0.05$	$\bar{a} = 0.10$	$\bar{\beta} = \widehat{\beta}_r$	$\bar{\beta} = \widehat{\beta}_u$	Iter.
Intercept	.6983 **	.4260 *	.7009 *	.5110 *	.5916 *	5456 *	.4729 *	.5050 *	.5124 *	.5136 *
	(.2608)	(.3140)	(.4029)	(.3266)	(.3162)	(.3167)	(.2973)	(.3060)	(.3066)	(.3059)
Age	.0165	.0374 *	.0232	.0354 *	.0302	0303	.0305 *	0312 *	.0320 *	.0320 *
	(.0302)	(.0314)	(.0345)	(.0318)	(.0318)	(.0310)	(.0298)	(.0310)	(.0314)	(.0313)
Age^2/10	-.0055	-.0045	-.0050	-.0047	-.0049	-.0045	-.0042	-.0046	-.0047	-.0047
	(.0090)	(.0092)	(.0091)	(.0092)	(.0091)	(.0088)	(.0086)	(.0090)	(.0092)	(.0091)
Female	-.1372	-.0697	-.1011	-.0691	-.0716	-.0724	-.0611	-.0834	-.0830	-.0833
	(.2454)	(.2532)	(.2551)	(.2549)	(.2538)	(.2479)	(.2399)	(.2505)	(.2529)	(.2518)
Female × Age	.0443 *	.0418 *	.0413 *	.0419 *	0413 *	.0388 *	0369 *	.0418 *	.0420 *	.0421 *
	(.0378)	(.0384)	(.0382)	(.0384)	(.0381)	(.0371)	(.0360)	(.0379)	(.0384)	(.0382)
Female × Age^2/10	-.0145 *	-.0163 *	-.0144 *	-.0160 *	-.0153 *	-.0145 *	-.0140 *	-.0156 *	-.0157 *	-.0157 *
	(.0115)	(.0118)	(.0119)	(.0118)	(.0118)	(.0115)	(.0111)	(.0116)	(.0118)	(.0117)
Couple	-.2141 *	.0756	-.1569	0341	-.0435	-.0205	0180	-.0067	-.0030	-.0034
	(.1162)	(.1733)	(.2916)	(.1914)	(.2099)	(.2025)	(.1834)	(.1711)	(.1732)	(.1724)
High Education	.4074 **	2710 **	3271 **	.2815 **	.2985 **	.2701 **	2527 **	2979 **	.3003 **	.3007 **
	(.1095)	(.1160)	(.1312)	(.1178)	(.1183)	(.1141)	(.1105)	(.1153)	(.1161)	(.1157)
Big City	-.2261 **	-.2063 *	-.2124 **	-.2123 *	-.2099 *	-.1983 *	-.1873 *	-.2070 *	-.2105 *	-.2109 *
	(.1041)	(.1073)	(.1058)	(.1067)	(.1052)	(.1028)	(.0998)	(.1055)	(.1067)	(.1062)
Employed	-.0893	-.1311	-.0893	-.1156	-.0989	-.0999	-.0981	-.1142	-.1153	-.1155
	(.1562)	(.1600)	(.1641)	(.1611)	(.1587)	(.1550)	(.1502)	(.1584)	(.1598)	(.1591)

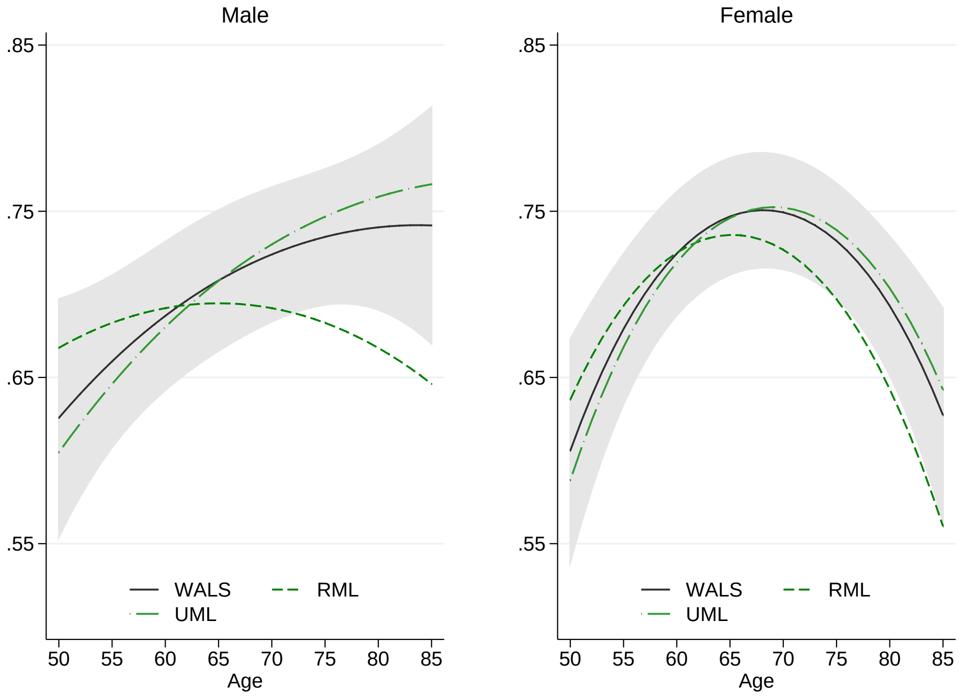
Notes: Sample size is 1,822 individuals and there are 256 possible models. $\hat{\beta}_r$ and $\hat{\beta}_u$ denote, respectively, the ML estimates based on the restricted and unrestricted models. FMA estimates are based on the smoothed AIC and BIC weighting systems. BMA estimates with conjugate priors for GLMs are based on the prior parameters $\bar{y} = 0.5\iota_n$ and $\bar{a} = \{0.01, 0.05, 0.10\}$. One-step WALS estimates with starting values $\bar{\beta} = \hat{\beta}_r$ and $\bar{\beta} = \hat{\beta}_u$ and iterative WALS estimates are based on the reflected Weibull prior. * denotes a t-ratio between 1 and 2, ** denotes a t-ratio greater than 2.

Table 3: Results of the Monte Carlo simulations

			ML		FMA			BMA			WALS		
Parameter	Criterion	n	$\widehat{\pi}_r$	$\widehat{\pi}_u$	BIC	AIC	FIC	$\bar{a} = 0.01$	$\bar{a} = 0.05$	$\bar{a} = 0.10$	$\bar{\beta} = \widehat{\beta}_r$	$\bar{\beta} = \widehat{\beta}_u$	Iter.
π_m	bias	100	0.1736	0.0112	0.0416	0.0249	0.0607	0.0583	0.0493	0.0619	0.1011	0.0473	0.0481
		400	0.0864	0.0019	0.0098	0.0007	0.0156	0.0112	0.0153	0.0268	0.0366	0.0165	0.0156
		900	0.0557	0.0018	0.0055	0.0024	0.0091	0.0044	0.0111	0.0220	0.0192	0.0091	0.0086
		1600	0.0407	0.0021	0.0036	0.0025	0.0060	0.0022	0.0100	0.0206	0.0114	0.0055	0.0053
	SE	100	0.1820	0.2446	0.2210	0.2282	0.2091	0.2203	0.2085	0.1948	0.1841	0.2190	0.2191
		400	0.0778	0.0934	0.0932	0.0925	0.0860	0.0926	0.0887	0.0844	0.0833	0.0872	0.0875
		900	0.0490	0.0596	0.0595	0.0590	0.0544	0.0591	0.0567	0.0541	0.0548	0.0558	0.0559
		1600	0.0370	0.0452	0.0451	0.0447	0.0411	0.0447	0.0429	0.0410	0.0420	0.0425	0.0425
	RMSE	100	0.2516	0.2448	0.2249	0.2295	0.2178	0.2279	0.2142	0.2044	0.2101	0.2241	0.2243
		400	0.1163	0.0934	0.0937	0.0925	0.0874	0.0933	0.0900	0.0886	0.0910	0.0888	0.0888
		900	0.0742	0.0597	0.0597	0.0590	0.0551	0.0593	0.0577	0.0584	0.0581	0.0566	0.0565
		1600	0.0550	0.0452	0.0452	0.0447	0.0416	0.0448	0.0440	0.0459	0.0435	0.0428	0.0428
π_f	bias	100	0.1302	0.0013	0.0140	0.0040	0.0336	0.0255	0.0278	0.0446	0.0811	0.0287	0.0275
		400	0.0615	0.0028	0.0006	0.0052	0.0085	0.0012	0.0098	0.0232	0.0297	0.0104	0.0093
		900	0.0383	0.0009	0.0015	0.0041	0.0054	0.0008	0.0090	0.0213	0.0155	0.0059	0.0054
		1600	0.0278	0.0005	0.0013	0.0028	0.0041	0.0005	0.0096	0.0215	0.0095	0.0040	0.0037
	SE	100	0.1323	0.2150	0.1786	0.1919	0.1634	0.1742	0.1710	0.1603	0.1497	0.1833	0.1818
		400	0.0589	0.0811	0.0776	0.0783	0.0696	0.0771	0.0750	0.0721	0.0704	0.0732	0.0732
		900	0.0368	0.0531	0.0505	0.0514	0.0455	0.0506	0.0493	0.0474	0.0471	0.0479	0.0479
		1600	0.0276	0.0403	0.0374	0.0389	0.0344	0.0380	0.0372	0.0358	0.0360	0.0363	0.0363
	RMSE	100	0.1857	0.2150	0.1792	0.1920	0.1668	0.1760	0.1732	0.1664	0.1702	0.1855	0.1838
		400	0.0852	0.0811	0.0776	0.0785	0.0701	0.0771	0.0757	0.0757	0.0765	0.0739	0.0738
		900	0.0531	0.0531	0.0505	0.0515	0.0458	0.0506	0.0501	0.0520	0.0496	0.0483	0.0483
		1600	0.0392	0.0403	0.0374	0.0390	0.0347	0.0380	0.0384	0.0418	0.0372	0.0365	0.0365

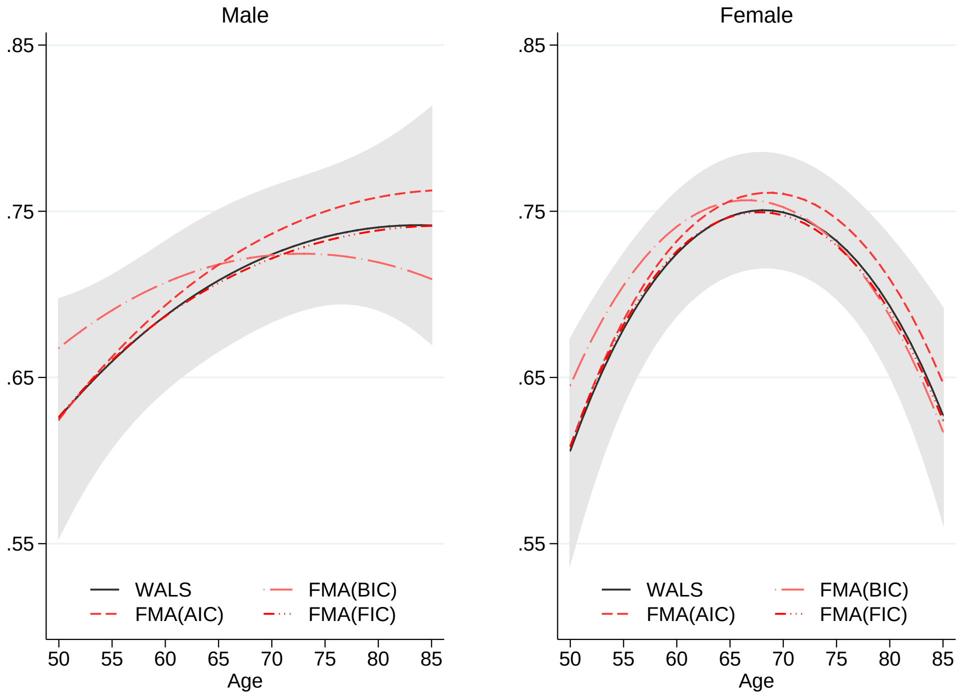
Notes: π_m and π_f denote, respectively, the participation probabilities of a representative male and a representative female aged 70 years. For all simulation designs, the true values of these parameters are equal to $\pi_m = 0.7301$ and $\pi_f = 0.7522$. $\widehat{\pi}_r$ and $\widehat{\pi}_u$ denote, respectively, the plug-in ML estimators of π_m and π_f in the restricted and unrestricted models. FMA estimators are based on the smoothed AIC, BIC and FIC weighting systems. BMA estimators with conjugate priors for GLMs are based on the prior parameters $\bar{y} = 0.5\iota_n$ and $\bar{a} = \{0.01, 0.05, 0.10\}$. One-step WALS estimators with starting values $\bar{\beta} = \hat{\beta}_r$ and $\bar{\beta} = \hat{\beta}_u$ and the iterative WALS estimator are based on the reflected Weibull prior. Monte Carlo results are computed by $1,000$ replications for each simulation design.

Figure 1: Iterative WALS and ML estimates of the participation probability age-profiles for representative male and female



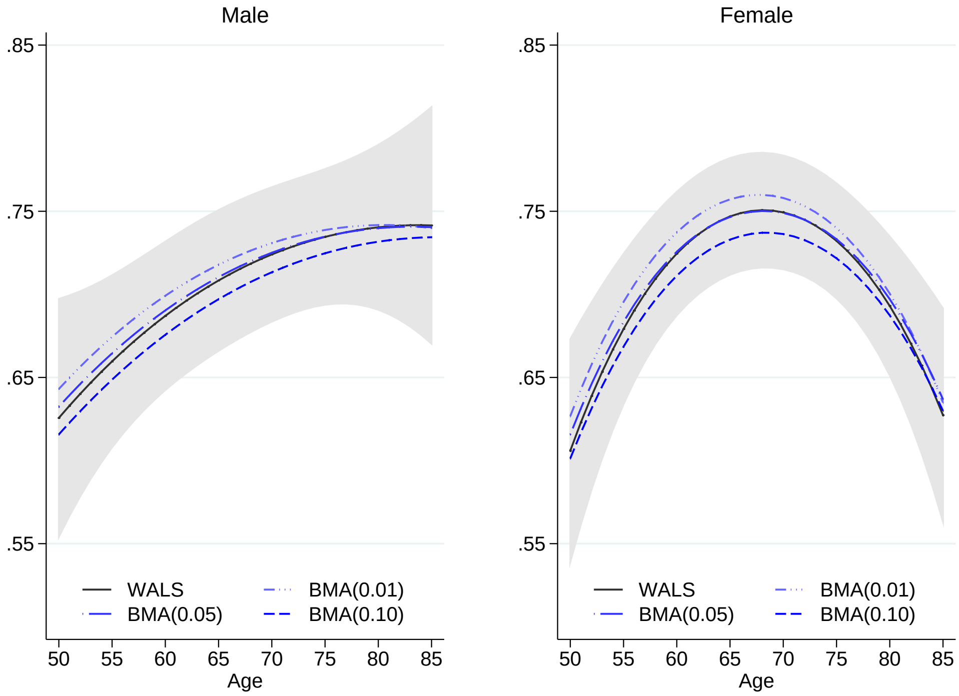
Notes: RML and UML denote, respectively, the plug-in ML estimates of π_{ma} and π_{fa} based on the restricted and unrestricted models, while WALS denotes the plug-in iterative WALS estimates (same as Figures 2 and 3). The shadow area is the one-std bands of the WALS estimates with standard errors computed by the delta method.

Figure 2: Iterative WALS and FMA estimates of the participation probability age-profiles for representative male and female



Notes: FMA(xIC) denotes the FMA estimates of π_{ma} and π_{fa} based on the smooth xIC (BIC, AIC, FIC) weighting system, while WALS denotes the plug-in iterative WALS estimates (same as Figures 1 and 3). The shadow area is the one-std bands of the WALS estimates with standard errors computed by the delta method.

Figure 3: Iterative WALS and BMA estimates of the participation probability age-profiles for representative male and female



Notes: BMA(x) denotes the BMA estimates of π_{ma} and π_{fa} based on the conjugate prior for GLMs with prior parameters $\bar{y} = 0.5\iota_n$ and $\bar{a} = x$, while WALS denotes the plug-in iterative WALS estimates (same as Figures 1 and 2). The shadow area is the one-std bands of the WALS estimates with standard errors computed by the delta method.

Appendix A: Proofs

Proof of Proposition 1. By the data transformations in (5), we can write the linearized system of constrained likelihood equations (4) for the jth model as

$$\begin{aligned} 0 &= \bar{X}_1'(\bar{y} - \bar{X}_1\beta_1 - \bar{X}_2\beta_2), \\ 0 &= \bar{X}_2'(\bar{y} - \bar{X}_1\beta_1 - \bar{X}_2\beta_2) - R_j\nu_j, \\ 0 &= R_j'\beta_2. \end{aligned} \tag{A1}$$

Given ν_j and ignoring the remainders in these approximations, the restricted one-step ML estimator $\widetilde{\beta}_j = (\widetilde{\beta}_{1j}', \widetilde{\beta}_{2j}')'$ solves the equation system

$$\begin{bmatrix} \bar{X}_1'\bar{X}_1 & \bar{X}_1'\bar{X}_2 \\ \bar{X}_2'\bar{X}_1 & \bar{X}_2'\bar{X}_2 \end{bmatrix} \begin{pmatrix} \widetilde{\beta}_{1j} \\ \widetilde{\beta}_{2j} \end{pmatrix} = \begin{pmatrix} \bar{X}_1'\bar{y} \\ \bar{X}_2'\bar{y} \end{pmatrix} - \begin{bmatrix} 0 \\ R_j \end{bmatrix} \nu_j,$$

while the unrestricted one-step ML estimator $\widetilde{\beta}_u = (\widetilde{\beta}_{1u}', \widetilde{\beta}_{2u}')'$ solves

$$\begin{bmatrix} \bar{X}_1'\bar{X}_1 & \bar{X}_1'\bar{X}_2 \\ \bar{X}_2'\bar{X}_1 & \bar{X}_2'\bar{X}_2 \end{bmatrix} \begin{pmatrix} \widetilde{\beta}_{1u} \\ \widetilde{\beta}_{2u} \end{pmatrix} = \begin{pmatrix} \bar{X}_1'\bar{y} \\ \bar{X}_2'\bar{y} \end{pmatrix}.$$

Rearranging these two expressions we obtain

$$\begin{pmatrix} \widetilde{\beta}_{1j} \\ \widetilde{\beta}_{2j} \end{pmatrix} = \begin{pmatrix} \widetilde{\beta}_{1u} \\ \widetilde{\beta}_{2u} \end{pmatrix} - \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \begin{bmatrix} 0 \\ R_j \end{bmatrix} \nu_j, \tag{A2}$$

where

$$\begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} = \begin{bmatrix} \bar{X}_1'\bar{X}_1 & \bar{X}_1'\bar{X}_2 \\ \bar{X}_2'\bar{X}_1 & \bar{X}_2'\bar{X}_2 \end{bmatrix}^{-1}.$$

Premultiplying both sides of (A2) by the $r_j \times k$ matrix $[0 : R_j']$ gives

$$[0 : R_j'] \begin{pmatrix} \widetilde{\beta}_{1j} \\ \widetilde{\beta}_{2j} \end{pmatrix} = [0 : R_j'] \begin{pmatrix} \widetilde{\beta}_{1u} \\ \widetilde{\beta}_{2u} \end{pmatrix} - [0 : R_j'] \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \begin{bmatrix} 0 \\ R_j \end{bmatrix} \nu_j.$$

Since $\widetilde{\beta}_{2j}$ satisfies the restriction $R_j'\widetilde{\beta}_{2j} = 0$ (by construction) and the matrix $R_j'\bar{A}_{22}R_j$ is nonsingular, solving this system of equations for the Lagrange multiplier gives

$$\widetilde{\nu}_j = (R_j'\bar{A}_{22}R_j)^{-1}R_j'\widetilde{\beta}_{2u}.$$

Thus, the restricted one-step ML estimators of β_1 and β_2 for the jth model can be written as

$$\widetilde{\beta}_{1j} = \widetilde{\beta}_{1u} - \bar{A}_{12}R_j(R_j'\bar{A}_{22}R_j)^{-1}R_j'\widetilde{\beta}_{2u}, \qquad \widetilde{\beta}_{2j} = \widetilde{\beta}_{2u} - \bar{A}_{22}R_j(R_j'\bar{A}_{22}R_j)^{-1}R_j'\widetilde{\beta}_{2u},$$

where $\bar{A}_{12} = -(\bar{X}_1'\bar{X}_1)^{-1}\bar{X}_1'\bar{X}_2(\bar{X}_2'\bar{M}_1\bar{X}_2)^{-1}$ and $\bar{A}_{22} = (\bar{X}_2'\bar{M}_1\bar{X}_2)^{-1}$, or equivalently

$$\widetilde{\beta}_{1j} = \widetilde{\beta}_{1u} + \bar{Q}\bar{P}_j\widetilde{\vartheta}, \qquad \widetilde{\beta}_{2j} = \widetilde{\beta}_{2u} - \left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{-1/2} \bar{P}_j\widetilde{\vartheta}.$$

The result then follows by noting that in the fully restricted model, where $R_j = I_{k_2}$ and $\bar{P}_j = I_{k_2}$, we obtain $\widetilde{\beta}_{1r} = \widetilde{\beta}_{1u} + \bar{Q}\widetilde{\vartheta} = (\bar{X}_1'\bar{X}_1)^{-1}\bar{X}_1'\bar{y}$. □

Proof of Proposition 2. Under the regularity conditions stated in the proposition, the one-step ML estimator for the unrestricted model has the same asymptotic distribution as the fully-iterated ML estimator and so $\sqrt{n}(\widetilde{\beta}_{un} - \beta_n) \Rightarrow \mathcal{N}(0, \Omega)$, where

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{bmatrix}^{-1} = \mathcal{I}^{-1},$$

with $\Omega_{11} = \mathcal{I}_{11}^{-1} + \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}\mathcal{I}_{21}\mathcal{I}_{11}^{-1}$, $\Omega_{12} = -\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}$, and $\Omega_{22} = \left(\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right)^{-1}$. Equation (6) also implies that

$$\sqrt{n}(\widetilde{\vartheta}_n - \vartheta_n) = \left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{1/2} \sqrt{n}(\widetilde{\beta}_{2un} - \beta_{2n}) + \left[\left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{1/2} - \Omega_{22}^{-1/2}\right]\delta,$$

with $\vartheta_n = \Omega_{22}^{-1/2}\beta_{2n}$. As $n \to \infty$, we have

$$\operatorname{plim}\left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{1/2} = \operatorname{plim}\left(\frac{\bar{H}_{22} - \bar{H}_{21}\bar{H}_{11}^{-1}\bar{H}_{12}}{n}\right)^{1/2} = \Omega_{22}^{-1/2}$$

and therefore

$$\sqrt{n}(\widetilde{\vartheta}_n - \vartheta_n) \Rightarrow \mathcal{N}(0, I_{k_2}). \tag{A3}$$

From Proposition 1 we have $\widetilde{\beta}_{1rn} = \widetilde{\beta}_{1un} + \bar{Q}\widetilde{\vartheta}_n$, or equivalently,

$$\sqrt{n}(\widetilde{\beta}_{1rn} - \beta_1) = \bar{Q}\Omega_{22}^{-1/2}\delta + \sqrt{n}(\widetilde{\beta}_{1un} - \beta_1) + \bar{Q}\sqrt{n}(\widetilde{\vartheta}_n - \theta_n).$$

Since $\operatorname{plim}\bar{Q} = \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\Omega_{22}^{1/2} = \mathcal{Q}$, we obtain

$$\sqrt{n}(\widetilde{\beta}_{1rn} - \beta_1) \Rightarrow \mathcal{N}(\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\,\delta, \mathcal{I}_{11}^{-1}). \tag{A4}$$

Moreover, $\widetilde{\beta}_{1rn}$ and $\widetilde{\vartheta}_n$ are asymptotically independent because their joint asymptotic distribution is normal with asymptotic covariance $\Omega_{12}\Omega_{22}^{-1/2} + \mathcal{Q} = 0$. For the one-step ML estimator of the jth model, Proposition 1 implies that

$$\sqrt{n}(\widetilde{\beta}_{1jn} - \beta_1) = \bar{Q}\bar{P}_j\Omega_{22}^{-1/2}\delta + \left[\sqrt{n}(\widetilde{\beta}_{1rn} - \beta_1) - \bar{Q}\Omega_{22}^{-1/2}\delta\right] - \bar{Q}\bar{W}_j\sqrt{n}(\widetilde{\vartheta}_n - \theta_n)$$

and

$$\sqrt{n}(\widetilde{\beta}_{2jn} - \beta_{2n}) = \left[\left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{-1/2} \bar{W}_j\Omega_{22}^{-1/2} - I_{k_2}\right]\delta + \left(\frac{\bar{X}_2'\bar{M}_1\bar{X}_2}{n}\right)^{-1/2} \bar{W}_j\sqrt{n}(\widetilde{\vartheta}_n - \theta_n).$$

The asymptotic distribution of $\widetilde{\beta}_{jn}$ then follows from (A3) and (A4), the asymptotic independence of $\widetilde{\beta}_{1rn}$ and $\widetilde{\vartheta}_n$, and the probability limits

$$\operatorname{plim}\bar{P}_j = \Omega_{22}^{1/2}R_j(R_j'\Omega_{22}R_j)^{-1}R_j'\Omega_{22}^{1/2} = \mathcal{P}_j, \qquad \operatorname{plim}\bar{W}_j = I_{k_2} - \mathcal{P}_j = \mathcal{W}_j.\square$$

Proof of Proposition 3. It follows from (11) and (13) that

$$\sqrt{n}(\widehat{\gamma}_n - \gamma_n) = \left(\begin{array}{c} \sqrt{n}(\widehat{\gamma}_{1n} - \gamma_1) \\ \sqrt{n}(\widehat{\gamma}_{2n} - \gamma_{2n}) \end{array} \right) = \left(\begin{array}{c} \sqrt{n}(\widetilde{\gamma}_{1rn} - \gamma_1) - \bar{D}W\sqrt{n}\widetilde{\gamma}_{2un} \\ W\sqrt{n}\widetilde{\gamma}_{2un} - d \end{array} \right),$$

where

$$\sqrt{n}(\widetilde{\gamma}_{1rn} - \gamma_1) \Rightarrow N_{1r} \sim \mathcal{N}(\mathcal{D}\,d, \mathcal{J}_{11}^{-1}), \qquad \sqrt{n}\widetilde{\gamma}_{2un} \Rightarrow N_{2u} \sim \mathcal{N}(d, I_{k_2}),$$

with $d = \sqrt{n}\gamma_{2n}$ and $W = W(N_{2u})$ because of (12). This implies that

$$\sqrt{n}(\widehat{\gamma}_n - \gamma_n) \Rightarrow N = \left(\begin{array}{c} N_1 \\ N_2 \end{array} \right) = \left(\begin{array}{c} N_{1r} - \mathcal{D}WN_{2u} \\ WN_{2u} - d \end{array} \right).$$

Moreover, since N_{1r} and N_{2u} are stochastically independent, we obtain

$$\mathbb{E}(N_1|N_{2u}) = \mathbb{E}(N_{1r}) - \mathcal{D}WN_{2u} = -\mathcal{D}(WN_{2u} - d)$$

and

$$\operatorname{var}(N_1|N_{2u}) = \operatorname{var}(N_{1r}) = \mathcal{J}_{11}^{-1}.$$

The asymptotic bias and the asymptotic variance of $\widehat{\gamma}_{1n}$ are equal, respectively, to the unconditional mean and the unconditional variance of the random vector N_1. The unconditional mean is given by

$$\mathrm{AB}(\widehat{\gamma}_{1n}) = \mathbb{E}[\mathbb{E}(N_1|N_{2u})] = -\mathcal{D}\,\mathbb{E}[\sqrt{n}(W\widetilde{\gamma}_{2un} - \gamma_{2n})] = -\mathcal{D}\,\mathbb{E}[\sqrt{n}(\widehat{\gamma}_{2n} - \gamma_{2n})] = -\mathcal{D}\,\mathrm{AB}(\widehat{\gamma}_{2n})$$

and the unconditional variance by

$$\mathrm{AV}(\widehat{\gamma}_{1n}) = \mathbb{E}[\operatorname{var}(N_1|N_{2u})] + \operatorname{var}[\mathbb{E}(N_1|N_{2u})] = \mathcal{J}_{11}^{-1} + \mathcal{D}\operatorname{var}(\sqrt{n}(\widehat{\gamma}_{2n} - \gamma_{2n}))\mathcal{D}' = \mathcal{J}_{11}^{-1} + \mathcal{D}\,\mathrm{AV}(\widehat{\gamma}_{2n})\mathcal{D}'.$$

The result for the AMSE follows. □

Appendix B: Continuity of eigenprojections and symmetric matrix functions

In matrix theory, when employing arguments that require limits such as continuity or consistency, some care is required when dealing with eigenvectors and associated concepts. Since there appears to be a certain amount of confusion on these issues among statisticians and econometricians, we present below some of the main results. Most of the results in this appendix are not new, see e.g. Kato (1976) and Horn and Johnson (1991, Chapter 6), but they are put together here in a simple and accessible manner in order to avoid further confusion.

Preliminaries

We shall confine ourselves to a real $n \times n$ symmetric matrix, say A. If $Ax = \lambda x$ for some $x \neq 0$ then λ is an eigenvalue of A and x is an eigenvector of A associated with λ. Because of the symmetry of A, all its eigenvalues are real and they are uniquely determined. However, eigenvectors are not uniquely determined, not even when the eigenvalue is simple. Also, while the eigenvalues are typically continuous functions of the elements of the matrix, this is not necessarily so for the eigenvectors. The current appendix attempts to make these vague notions precise.

Some definitions are required. The set of all eigenvalues of A is called its *spectrum* and is denoted as $\sigma(A)$. The *eigenspace* of A associated with λ is

$$V(\lambda) = \{x \in \mathbb{R}^n | Ax = \lambda x\}.$$

The dimension of $V(\lambda)$ is equal to the multiplicity of λ, say $m(\lambda)$. Eigenspaces associated with distinct eigenvalues are orthogonal to each other. Because of the symmetry of A we have the decomposition

$$\sum_{\lambda \in \sigma(A)} V(\lambda) = \mathbb{R}^n.$$

The *eigenprojection* of A associated with λ of multiplicity $m(\lambda)$, denoted $P(\lambda)$, is given by the symmetric idempotent matrix

$$P(\lambda) = \sum_{j=1}^{m(\lambda)} x_j x_j',$$

where the $\{x_j\}$ form any set of m orthonormal vectors in $V(\lambda)$, that is, $x_j' x_j = 1$ and $x_i' x_j = 0$ for $i \neq j$. While eigenvectors are not unique, the eigenprojection is unique because an idempotent matrix is uniquely determined by its range and null space. The spectral decomposition of A is then

$$\sum_{\lambda \in \sigma(A)} \lambda P(\lambda) = A.$$

If σ_0 is any subset of $\sigma(A)$, then the *total eigenprojection* associated with the eigenvalues in σ_0 is defined as

$$P(\sigma_0) = \sum_{\lambda \in \sigma_0} P(\lambda).$$

It is clear that $P(\sigma(A)) = I_n$. Also, if σ_0 contains only one eigenvalue, say λ, then $P(\{\lambda\}) = P(\lambda)$. Total eigenprojections are a key concept when dealing with limits, as we shall see below.

Symmetric matrix functions

Now consider a matrix function $A(t)$, where $A(t)$ is a real $n \times n$ symmetric matrix for every real t. The matrix $A(t)$ has n eigenvalues, say $\lambda_1(t), \dots, \lambda_n(t)$, some of which may be equal. Suppose that $A(t)$ is continuous at $t = 0$. Then the eigenvalues are also continuous at $t = 0$. This was proved by Rellich (1953) making essential use of the symmetry of $A(t)$.

Now, let λ be an eigenvalue of $A = A(0)$ of multiplicity m. Because of the continuity of the eigenvalues we can separate the eigenvalues in two groups, say $\lambda_1(t), \dots, \lambda_m(t)$ and $\lambda_{m+1}(t), \dots, \lambda_n(t)$, where the m eigenvalues in the first group converge to λ, while the $n - m$ eigenvalues in the second group also converge, but not to λ. Kato (1976, Theorem 5.1), based on earlier results by Rellich (1937, 1953), proved that the total eigenprojection $P(\{\lambda_1(t), \dots, \lambda_m(t)\})$ is continuous at $t = 0$, that is, it converges to the spectral projection $P(\lambda)$ of $A(0)$.

Kato's result does *not* imply that eigenvectors or eigenprojections are continuous. If all eigenvalues of $A(t)$ are distinct at $t = 0$ then each eigenprojection $P_j(t)$ is continuous at $t = 0$ because it coincides with the total eigenprojection for the eigenvalue $\lambda_j(t)$. But if there are multiple eigenvalues at $t = 0$, then it may occur that the eigenprojections do not converge as $t \to 0$, unless we assume that the matrix $A(t)$ is (real) analytic. (A function is real analytic if it is infinitely differentiable and can be expanded in a power series.) In fact, Kato (1976, Theorem 1.10) showed that if $A(t)$ is real analytic at $t = 0$, then the eigenvalues and the eigenprojections are also analytic at $t = 0$ (and therefore certainly continuous).

Discontinuity of eigenprojections

Hence, in general, eigenvalues are continuous, but eigenvectors and eigenprojections may not be. This is well illustrated by the following example of Kato (1976, Example 5.3), which is adapted from Rellich (1937).

Consider the matrix

$$A(t) = e^{-1/t^2} \begin{pmatrix} \cos(2/t) & \sin(2/t) \\ \sin(2/t) & -\cos(2/t) \end{pmatrix}, \qquad A(0) = 0.$$

There is a multiple eigenvalue 0 at $t = 0$ and simple eigenvalues $\lambda_1 = e^{-1/t^2}$ and $\lambda_2 = -e^{-1/t^2}$ at $t \neq 0$. The associated eigenvectors are

$$x_1 = \begin{pmatrix} \cos(1/t) \\ \sin(1/t) \end{pmatrix}, \qquad x_2 = \begin{pmatrix} \sin(1/t) \\ -\cos(1/t) \end{pmatrix}.$$

Hence the associated eigenprojections are

$$P_1(t) = x_1 x_1' = \begin{pmatrix} \cos^2(1/t) & \sin(1/t)\cos(1/t) \\ \sin(1/t)\cos(1/t) & \sin^2(1/t) \end{pmatrix}$$

and

$$P_2(t) = x_2 x_2' = \begin{pmatrix} \sin^2(1/t) & -\sin(1/t)\cos(1/t) \\ -\sin(1/t)\cos(1/t) & \cos^2(1/t) \end{pmatrix}.$$

The matrix function $A(t)$ is continuous (even infinitely differentiable) for all real t. This is also true for the eigenvalues. But there is no eigenvector which is continuous in the neighborhood of $t = 0$ and does not vanish at $t = 0$. Also, the eigenprojections $P_1(t)$ and $P_2(t)$, while continuous (even infinitely differentiable) in any interval not containing $t = 0$, cannot be extended to $t = 0$ as continuous functions.

The total eigenprojection is given by $P_1(t) + P_2(t) = I_2$, which is obviously continuous at $t = 0$, but the underlying eigenprojections $P_1(t)$ and $P_2(t)$ are not. The reason lies in the fact that the matrix $A(t)$, while infinitely differentiable at $t = 0$, is not analytic.

This can be seen as follows. Let

$$f(t) = \begin{cases} \exp(-1/t^2) & \text{for } t \neq 0 \\ 0 & \text{for } t = 0, \end{cases} \qquad g(t) = \begin{cases} \cos(2/t) & \text{for } t \neq 0 \\ 0 & \text{for } t = 0, \end{cases}$$

and define $h(t) = f(t)g(t)$. It is well-known (and a standard example in textbooks) that the function $f(t)$ is infinitely differentiable for all (real) t, but not analytic. The function $g(t)$ is not continuous at $t = 0$, although it is infinitely differentiable in any interval not containing $t = 0$. Their product $h(t)$ is infinitely differentiable for all (real) t (because g is bounded), but it is not analytic.

We summarize the previous discussion as follows.

Lemma B.1: Let $A(t)$ be a family of real-valued symmetric matrices, and suppose $\epsilon > 0$ exists such that $A(t)$ is continuous for all $|t| < \epsilon$. Then the eigenvalues $\lambda_j(t)$ and the total eigenprojections $P_j(t)$ are continuous at $t = 0$. If, in addition, $A(t)$ is analytic at $t = 0$, then the individual eigenprojections are continuous at $t = 0$.

Relation to Tyler's lemma

Tyler (1981, Lemma 2.1) stated the following result, which is often quoted, but is essentially the same as Kato's result. Let $A(t)$ be a symmetric $n \times n$ matrix function with eigenvalues

$$\lambda_1(t) \geq \lambda_2(t) \geq \cdots \geq \lambda_i(t) \geq \cdots \geq \lambda_j(t) \geq \cdots \geq \lambda_n(t),$$

and assume that, at $t = 0$,

$$\lambda_{i-1}(0) > \lambda_i(0), \qquad \lambda_j(0) > \lambda_{j+1}(0).$$

If $A(t)$ is continuous at $t = 0$, then the total eigenprojection $P_{i,j}(t)$ associated with $\lambda_i(t), \ldots, \lambda_j(t)$ is continuous at $t = 0$.

Continuity of symmetric matrix functions

We are now in a position to state the following result, which is essentially the same as Horn and Johnson (1991, Theorem 6.2.37) but with a somewhat simpler proof.

Lemma B.2: Let $A(t)$ be a family of real-valued symmetric matrices, and suppose $\epsilon > 0$ exists such that $A(t)$ is continuous for all $|t| < \epsilon$. Let f be a real-valued function, defined and continuous on the spectrum $\sigma(A(0))$. Then $f(A(t))$ converges to $f(A(0))$ as $t \to 0$.

Proof: Since $A(t)$ is symmetric and continuous in t, we can write

$$A(t) = \sum_{\lambda(t) \in \sigma(A(t))} \lambda(t) P(\lambda(t)).$$

Let λ_0 be an eigenvalue of $A(0)$, and let

$$\lambda_i(t) \geq \cdots \geq \lambda_j(t) \qquad (0 < |t| < \epsilon)$$

be the λ-group associated with λ_0. Then,

$$\lim_{t \to 0} \lambda_k(t) = \lambda_0 \qquad (i \leq k \leq j),$$

and hence, since f is continuous at λ_0,

$$\lim_{t \to 0} f(\lambda_k(t)) = f(\lambda_0) \qquad (i \leq k \leq j).$$

We also know, because of the continuity of the total eigenprojections, that

$$\lim_{t\to 0}\sum_{k=i}^{j} P(\lambda_k(t)) = P(\lambda_0).$$

Together this implies that

$$\lim_{t\to 0}\sum_{k=i}^{j} f(\lambda_k(t))P(\lambda_k(t)) = f(\lambda_0)P(\lambda_0),$$

which we see by writing

$$\begin{aligned}&\sum_{k=i}^{j} f(\lambda_k(t))P(\lambda_k(t)) - f(\lambda_0)P(\lambda_0)\\ &\qquad = \sum_{k=i}^{j}\big[f(\lambda_k(t)) - f(\lambda_0)\big]P(\lambda_k(t)) - f(\lambda_0)\big[(P(\lambda_0) - \sum_{k=i}^{j} P(\lambda_k(t))\big].\end{aligned}$$

This proves convergence for each λ-group, and hence concludes the proof.

Orthogonal transformations

Let B be an $m \times n$ matrix of full column-rank n. Then $A = B'B$ is positive definite and symmetric, and we can decompose

$$A = T\Lambda T',$$

where Λ is diagonal with strictly positive elements and T is orthogonal.

Suppose that our calculations would be much simplified if A were equal to the identity matrix. We can achieve this by transforming B to a matrix C, as follows:

$$C = BT\Lambda^{-1/2}S',$$

where S is an arbitrary orthogonal matrix. Then,

$$C'C = S\Lambda^{-1/2}T'B'BT\Lambda^{-1/2}S' = S\Lambda^{-1/2}\Lambda\Lambda^{-1/2}S' = SS' = I_n.$$

The matrix S is completely arbitrary, as long as it is orthogonal. It is tempting to choose $S = I_n$. This, however, implies that if $B = B(t)$ is a continuous function of some variable t, then $C = C(t)$ is *not* necessarily continuous, as is shown by the previous discussion. There is only one choice of S that leads to continuity of C, namely $S = T$, in which case

$$C = BT\Lambda^{-1/2}T' = B(B'B)^{-1/2}.$$

References

Ando, T., and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* 109: 254–265.

Ando, T., and Li, K.-C. (2017). A weight-relaxed model averaging approach for high dimensional generalized linear models. *Annals of Statistics*, forthcoming.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* 85: 233–298.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics* 41: 802–837.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics* 53: 603–618.

Chen, M. H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis* 3: 585–614.

Chen, M. H., and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica* 13: 461–476.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review: Papers & Proceedings* 105: 486–490.

Claeskens, G., Croux, C., and van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* 62: 972–979.

Claeskens, G., and Hjort, N. L. (2003). The focused information criterion (with discussion). *Journal of the American Statistical Association* 98: 900–916.

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

Clyde, M. A. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics* 11: 745–763.

Clyde, M. A., and George, E. I. (2004). Model uncertainty. *Statistical Science* 19: 81–94.

Danilov, D., and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics* 122: 27–46.

De Luca, G., and Magnus, J. R. (2011). Bayesian model averaging and weighted average least squares: Equivariance, stability and numerical issues. *Stata Journal* 11: 518–544.

Fahrmeir, L., and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics* 13: 342–368.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348–1360.

Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20: 101–148.

Hansen, B. E. (2014). Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics* 5: 495–530.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics* 190: 115–132.

Heumann, C., and Grenke, M. (2010). An efficient model averaging procedure for logistic regression models using a Bayesian estimator with Laplace prior. In T. Kneib and G. Tutz (eds) *Statistical Modelling and Regression Structures.* Physica-Verlag, Heidelberg, pp. 79–90.

Hjort, N. L., and Claeskens, G. (2003a). Frequentist model averaging estimators (with discussion). *Journal of the American Statistical Association* 98: 879–899.

Hjort, N. L., and Claeskens, G. (2003b). Rejoinder to "Frequentist model average estimators" and "The focused information criterion". *Journal of the American Statistical Association* 98: 938–945.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science* 14: 382–417.

Horn, R. A., and Johnson, C. R. (1991). *Topics in Matrix Analysis.* Cambridge University Press, Cambridge/ New York.

Ishwaran H., and Rao J. S. (2003). Discussion to "Frequentist model average estimators" and "The focussed information criterion" by Hjort, N. L. and Claeskens, G. *Journal of the American Statistical Association* 98: 922–925.

Kato, T. (1976). *Perturbation Theory for Linear Operators*, 2nd edition. Springer-Verlag, Berlin/ Heidelberg/ New York.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, Cambridge.

Kumar, K., and Magnus, J. R. (2013). A characterization of Bayesian robustness for a normal location parameter. *Sankhya (Series B)* 75: 216–237.

Leeb, H., and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* 19: 100–142.

Leeb, H., and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34: 2554–2591.

Liu, C. A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics* 186: 142–159.

Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications* 44: 293–308.

Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with known variance. *Econometrics Journal* 5: 225–236.

Magnus, J. R., and De Luca, G. (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys* 30: 117–148.

Magnus, J. R., and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67: 639–643.

Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two averaging techniques with an application to growth empirics. *Journal of Econometrics* 154: 139–153.

Malter, F., and Börsch-Supan, A. (2015). *SHARE Wave 5: Innovations & Methodology.* MEA, Max Planck Institute for Social Law and Social Policy, Munich.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys* 29: 46–75.

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society (Series A)* 135: 370–384.

Newey, N. K., and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (eds), *Handbook of Econometrics*, Vol. 4. North-Holland, Amsterdam, pp. 2111–2245.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 103: 681–686.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* 83: 251–266.

Raftery, A. E., and Zheng, Y. (2003). Discussion to "Frequentist model average estimators" and "The focussed information criterion" by Hjort, N. L. and Claeskens, G. *Journal of the American Statistical Association* 98: 931–938.

Rellich, F. (1937). Störungstheorie der Spektralzerlegung. *Mathematische Annalen*, 113, 600–619.

Rellich, F. (1953, 1969). *Perturbation Theory of Eigenvalue Problems.* Gordon & Breach, New York.

Robinson, P.M. (1988). The stochastic difference between econometric statistics. *Econometrica* 56: 531–548.

Tyler, D. E. (1981). Asymptotic inference for eigenvalues. *The Annals of Statistics*, 9, 725–736.

Zou, G., Wan, A. T. K., Wu, X., and Chen, T. (2007). Estimation of regression coefficients of interest when other regression coefficients are of no interest: The case of non-normal errors. *Statistics & Probability Letters* 77: 803–810.