# OLS with multiple high dimensional category variables

By

Gaure, Simen

# OLS with Multiple High Dimensional Category Variables

Simen Gaure[1,2]

**Abstract**

A new algorithm is proposed for OLS estimation of linear models with multiple high-dimensional category variables. It is a generalization of the within transformation to arbitrary number of category variables. The approach, unlike other fast methods for solving such problems, provides a covariance matrix for the remaining coefficients. The article also sets out a method for solving the resulting sparse system, and the new scheme is shown, by some examples, to be comparable in computational efficiency to other fast methods. The method is also useful for transforming away groups of pure control dummies. A parallelized implementation of the proposed method has been made available as an R-package `lfe` on CRAN.

*Keywords:* Alternating Projections, Kaczmarz Method, Two-Way Fixed Effects, Multiple Fixed Effects, High Dimensional Category Variables, Panel Data

## 1. Introduction

We consider OLS estimation of models of the form

$$y = X\beta + D\alpha + \epsilon, \tag{1}$$

*Email address:* `Simen.Gaure@frisch.uio.no` (Simen Gaure)
*URL:* `http://www.frisch.uio.no` (Simen Gaure)
[1]Ragnar Frisch Centre for Economic Research, Gaustadalléen 21, N-0349 Oslo, Norway. Tel: +47 22958817
[2]Research Computing Services, USIT, University of Oslo

where $y$ is a response vector of length $n$, $X$ is an $n \times k$-matrix of covariates with corresponding parameters $\beta$, $D$ is an $n \times g$-matrix of dummies with corresponding parameters $\alpha$, and $\epsilon$ is a normally distributed stochastic term. $D$ is assumed to arise from dummy-encoding of one *or more* category variables, and it is assumed that each of these $e$ category variables have a large number of different values (large as in $10^5$–$10^7$).

That is, $D$ is a block matrix, $D = \begin{bmatrix} D_1 & D_2 & \cdots & D_e \end{bmatrix}$. The entries of each $D_i$ consists of 0 and 1, with 1 non-zero entry per row. Hence, the columns of each $D_i$ are pairwise orthogonal. In general, however, $D_i$ is not orthogonal to $D_j$ for $i \neq j$. We also assume that $k$ is reasonably small, and that the system without dummies is therefore manageable.

Models like this have been used to analyze panel data in the econometrics literature, e.g. in Abowd et al. (1999), which studies wage as an outcome $y$, and where a dummy variable is introduced for each individual ($D_1$), as well as for each firm ($D_2$). As the individuals move between different firms, it is not a nested model. The dummy modelling is done to account for arbitrarily distributed time-constant unobserved heterogeneity, both among employees and among firms. In this setting, what is studied is the correlation between the firm effect and the individual effect on the wage. Other applications of similar models can be found in (Aaronson and Barrow, 2007; Abowd et al., 2006; Bazzoli et al., 2008; Carneiro et al., 2012; Cornelißen and Hübler, 2011; Gibbons et al., 2010; Jacob and Lefgren, 2008), e.g. with school and teacher effects, worker and area effects, or hospital and doctor effects. Because the datasets, sourced from government registers containing hundreds of thousands, or even millions, of observations, are so large, there is a very large number of dummies, requiring special algorithms for their estimation.

McCaffrey et al. (2010) compares some estimation schemes for these models. We present a new estimation algorithm for such models. In an appendix, we present some test runs on both simulated and real datasets, comparing the suggested method to `a2reg` by Ouazad (2008), the fastest one in McCaffrey et al. (2010). Our method has the following properties.

- For the system (1), $\hat{\beta}$ is found without having to find $\hat{\alpha}$, i.e. large matrices are avoided.

- Since the method is a direct generalization of the within-groups estimator, utilizing the Frisch-Waugh-Lovell theorem, it yields a covariance matrix for $\hat{\beta}$, which in the case of $e = 2$ is the same as if (1) had been solved directly with standard OLS. In the case $e > 2$, the covariance matrix may be slightly up-scaled. The other methods of McCaffrey et al. (2010) lack this property: either no standard errors are provided, or they need to be computed for both $\hat{\alpha}$ and $\hat{\beta}$ requiring the use of various time-consuming methods. In the approach proposed here, however, it is possible to have a number of pure control dummies in $D$ which are simply projected out of the system and not estimated, and we still get good estimates for $\hat{\beta}$. An example can be found in Markussen and Røed (2012).

- $\hat{\alpha}$ may optionally be retrieved, but identification of coefficients in $\hat{\alpha}$ in the case $e > 2$ hinges on the ability of the researcher to find enough estimable functions suitable for the problem at hand. The estimation scheme provides a test for estimability. Some examples are given in the appendix.

- The method is comparable in speed to `a2reg`, sometimes faster and sometimes slower.

In Andrews et al. (2008) it is shown that the above-mentioned correlation is negatively biased; the present approach does not concern itself with the appropriateness of those models to the solution of particular problems, nor with their statistical properties. What it does concern itself with, however, is the OLS estimation of the parameter vectors $\beta$ and $\alpha$ only.

Although we occasionally refer to the firm-employee application mentioned above, it is only because it gives us convenient names for the category variables. While the main application of our results is in the analysis of panel data, the results are general in the sense that they do not depend on any particular appli-

cation; any set of dummies $D$, be it age-groups, hometown, school, workplace, may be transformed out of equation (1). The approach does not even depend on whether the underlying data are panel data; we are not concerned with time-constant and time-varying covariates, nor, indeed, time in general, just with sets of possibly high-dimensional category variables in OLS estimation.

**Remark 1.1.** In the econometrics literature, having a time-constant dummy for each individual is referred to as having *individual fixed effects.* Thus, to (some) econometricians, the $\alpha$ above is *the fixed effects,* whereas $\beta$ is not. To statisticians, this use of the phrase "fixed effect" is confusing, as the $\beta$ is also a fixed effect. Below, we use the phrase "fixed effect" in the econometrician's sense. I apologize to statisticians who might (rightfully) find this confusing, but it is done to (hopefully) make the text more accessible to econometricians.

A common strategy when working with a single category variable, is to centre the covariates and response on the group means, and do OLS on the projected system (Wooldridge, 2002, Section 10.5). Such centring consists of computing, for each category, e.g. individual, the mean of the covariate. The mean is then subtracted from the covariate's values in the category. Thus, time-constant effects are removed, and only time-varying effects *within* each category matter in the estimation. That is, the category effect serves to replace all time-constant effects.

Centring on the means is also referred to as "demeaning", "time demeaning", "fixed effects transformation", "within transformation", "within groups transformation" or "sweeping out the fixed effects". It seems to be common knowledge that sweeping out more than one category variable may not be done by centring on the group means, or by other simple transformations of the data, see e.g. (Abowd et al., 1999, p. 266), (Andrews et al., 2008, p. 676), (Cornelißen and Hübler, 2011, p. 476), and (Margolis and Salvanes, 2001, p. 19). Other estimation methods have therefore been developed to meet this challenge. Several authors, (e.g., Abowd et al., 2002; Cornelißen, 2008; Ouazad, 2008; Schmieder, 2009), have implemented procedures for the estimation of such models.

The main contribution of this work is Remark 3.2. It is indeed possible to sweep out multiple fixed effects, due to the Frisch-Waugh-Lovell theorem and certain other, relatively old results (von Neumann, 1949, Lemma 22, p.475), and (Halperin, 1962, Theorem 1)[3], now known as *The Method of Alternating Projections*. This leaves us with two systems, one of which is manageable with off-the-shelf OLS software; the other, a large, sparse system for the fixed effects. As a bonus we get a covariance matrix for the small system, the $\beta$s, a property which is lacking in many of the other estimation schemes. In Algorithm 6.1 and the discussion preceding it, we also suggest a method, the Kaczmarz method (Kaczmarz, 1937), for solving the sparse system.

To the author's knowledge, both the Kaczmarz method and Halperin's method of alternating projections are little known in the econometric and statistical literature, even though they sit very nicely with the Frisch-Waugh-Lovell theorem. There is an application of Halperin's theorem to the proof of convergence of the backfitting algorithm in (Ansley and Kohn, 1994), and the Kaczmarz method is actively in use in medical imaging, under the name "Algebraic Reconstruction Technique" (ART), (Andersen and Kak, 1984; Gordon et al., 1970; Herman, 2009; Hudson and Larkin, 1994).

To get an intuitive graphical view of these methods, and why they matter when studying linear systems, recall that a line, a plane, hyperplane or linear subspace is merely the solution set of one or more linear equations. The intersection of such subspaces is the simultaneous solution of the corresponding sets of equations. Consider e.g. two intersecting lines in Euclidean plane. To find the intersection, we can start at a point anywhere in the plane, project it orthogonally onto one of the lines, project it again onto the other line, then back to the first line, and so on. We zigzag towards the intersection. Clearly, if the lines are orthogonal, we will reach the intersection in just two steps. If the lines intersect at an acute angle, more steps will be needed to get close. The Kaczmarz method is the generalization of this process to a finite set of

[3]Halperin's article is available here: http://bit.ly/HJ067o

hyperplanes in a high dimensional Euclidean space. von Neumann's lemma is the generalization to two arbitrary subspaces of a Banach space. Halperin's theorem generalizes this further, to a finite number of subspaces.

The method presented falls in the broad category of sparse methods. Other recent advances in sparse methods may be found in Lee and Huang (2013); Vidaurre et al. (2013), whereas a method for choosing between different models may be found in Ueki and Kawasaki (2013).

## 2. Preliminaries

In (1), we assume that the model is well specified in the sense that the only multicollinearities in the system occur in $D$. Some of these multicollinearities are due to the fact that we have constructed $D$ from a full set of dummies, with no references. Others may be due to spurious relations between the category variables. We return to the mathematical details, once we have introduced some notation.

The task is to compute the OLS estimates $\hat{\beta}$ and $\hat{\alpha}$ of the parameter vectors $\beta$ and $\alpha$ in (1). In particular we look at the case $e = 2$, corresponding to two category variables, e.g. "firm" and "employee" as in (Abowd et al., 1999; Andrews et al., 2008).

We now derive the Frisch-Waugh-Lovell theorem. To do this, we consider for the time being a full-rank version $\mathcal{D}$ of $D$. I.e. we remove just enough linearly dependent columns from $D$ to get a new full-rank matrix $\mathcal{D}$ with the same range and with $\operatorname{rank}(\mathcal{D}) = \operatorname{rank}(D)$. That is, $\mathcal{D}'\mathcal{D}$ is invertible. The manner in which linearly dependent columns are removed is not important for most of the results of this section, since the results are in terms of the range projection of $D$, which only depends on the column space. We also remove the corresponding coordinates from $\alpha$ to get an $\alpha_r$.

We will have occasion to use several of the intermediate formulae later. An identity matrix of the appropriate size will generally be denoted by $I$. The transpose of any matrix $M$ is denoted by $M'$.

The normal equations of system (1), with $\mathcal{D}$ and $\alpha_r$ are

$$\begin{bmatrix} X'X & X'\mathcal{D} \\ \mathcal{D}'X & \mathcal{D}'\mathcal{D} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha}_r \end{bmatrix} = \begin{bmatrix} X' \\ \mathcal{D}' \end{bmatrix} y. \tag{2}$$

We recall some standard facts about these. We may write them as two rows

$$X'X\hat{\beta} + X'\mathcal{D}\hat{\alpha}_r = X'y \tag{3}$$

$$\mathcal{D}'X\hat{\beta} + \mathcal{D}'\mathcal{D}\hat{\alpha}_r = \mathcal{D}'y \tag{4}$$

and do Gaussian elimination of the $\hat{\alpha}_r$-term in the last row to get

$$X'(I - \mathcal{D}(\mathcal{D}'\mathcal{D})^{-1}\mathcal{D}')X\hat{\beta} = X'(I - \mathcal{D}(\mathcal{D}'\mathcal{D})^{-1}\mathcal{D}')y. \tag{5}$$

Now, let $P = I - \mathcal{D}(\mathcal{D}'\mathcal{D})^{-1}\mathcal{D}'$ and note that $P = P^2 = P'$ is a projection. Indeed, $P$ is the projection on the orthogonal complement of the column space of $\mathcal{D}$. Rewriting equation (5) with $P$, we get

$$(PX)'(PX)\hat{\beta} = (PX)'Py,$$

which shows that $\hat{\beta}$ is the OLS solution of the projected system

$$Py = PX\beta + P\epsilon. \tag{6}$$

We do not need $\hat{\alpha}$, then, to find $\hat{\beta}$. Moreover, by multiplying through (4) with $\mathcal{D}(\mathcal{D}'\mathcal{D})^{-1}$ and noting that $\mathcal{D}(\mathcal{D}'\mathcal{D})^{-1}\mathcal{D}' = I - P$, we get

$$(I - P)X\hat{\beta} + \mathcal{D}\hat{\alpha}_r = (I - P)y, \tag{7}$$

which may be reordered as

$$y - (X\hat{\beta} + \mathcal{D}\hat{\alpha}_r) = Py - PX\hat{\beta},$$

showing that the residuals of the projected system (6) are the same as the residuals of the full system (1).

In practice, it sometimes happens that $\hat{\beta}$ is not uniquely determined by (6). This is the same problem that affects ordinary within-groups models with $e = 1$, e.g. covariates in $X$ which are constant for individuals. We do not treat this

specification problem here. That is, we assume $PX$ is full rank. Similarly, we assume that $Q\mathcal{D}$ is full rank, where $Q$ is the projection onto the orthogonal complement of the range of $X$. That is, loosely speaking, $X$ should not explain any category in $D$ fully, nor should $D$ explain any of the variables in $X$ fully. In other words, barring those we have temporarily removed from $D$, the system should be free of other multicollinearities.

Now, recall that to find the covariance matrix of $\hat{\beta}$ and $\hat{\alpha}$ in system (1), we invert the matrix in (2) and multiply with the residual sum of squares divided by the degrees of freedom. Also, recall from general matrix theory the block inversion formula

$$\begin{bmatrix} A & B \\ C & E \end{bmatrix}^{-1} = \begin{bmatrix} (A - BE^{-1}C)^{-1} & -A^{-1}B(E - CA^{-1}B)^{-1} \\ -E^{-1}C(A - BE^{-1}C)^{-1} & (E - CA^{-1}B)^{-1} \end{bmatrix},$$

which is easily shown to hold when the quantities in question are invertible. We use this on the matrix in (2). After some simplification, we get

$$\begin{bmatrix} X'X & X'\mathcal{D} \\ \mathcal{D}'X & \mathcal{D}'\mathcal{D} \end{bmatrix}^{-1} = \begin{bmatrix} (X'PX)^{-1} & -(X'X)^{-1}X'\mathcal{D}(\mathcal{D}'Q\mathcal{D})^{-1} \\ -(\mathcal{D}'\mathcal{D})^{-1}\mathcal{D}'X(X'PX)^{-1} & (\mathcal{D}'Q\mathcal{D})^{-1} \end{bmatrix},$$

so that the upper left entry of the inverse, i.e. the part that goes into the covariance matrix for $\hat{\beta}$, is $(X'PX)^{-1}$. Now, $(X'PX)^{-1}$ is also the inverse of the corresponding matrix in system (6).

**Remark 2.1.** Getting the same matrix, and the same residuals, we can perform an OLS on system (6) in the normal way, adjusting only the degrees of freedom according to the number of parameters projected out by $P$, to find the $\hat{\beta}$ part of the covariance matrix for system (1).

The above result is known as the Frisch-Waugh-Lovell theorem. It is a standard way to eliminate the fixed effects from the equation in the case $e = 1$, i.e. with a single fixed effect. In this case, the projection $P$ is the *within transformation*.

**Remark 2.2.** Note that $P$ is the projection onto $\mathrm{R}(\mathcal{D})^{\perp}$, where $\mathrm{R}(\mathcal{D})$ denotes the range of $\mathcal{D}$, i.e. the column space, and $^{\perp}$ denotes orthogonal complement.

We have $\mathrm{R}(\mathcal{D}) = \mathrm{R}(D)$ since we have only removed linearly dependent columns, so $P$ and $\hat{\beta}$ do not depend on the specifics of how we produce $\mathcal{D}$ from $D$. Nor does the matrix $(X'PX)^{-1}$.

## 3. Projecting the system

Although (6) shows us how to eliminate the fixed effects for arbitrary $e \geq 1$, and we even have an explicit matrix formula for $P$, due to size limitations with large $n \approx 10^7$, it is impractical to handle the $n \times n$ projection matrix $P$ directly. However, we do not really need to find the matrix $P$; what we do need to do is compute $Py$ and $PX$.

To this end, for each $i = 1..e$ let $P_i$ be the projection onto the orthogonal complement of the range of $D_i$, $\mathrm{R}(D_i)^\perp$. $P_i$ is the within-groups transformation for category variable $i$. We have

$$\mathrm{R}(D)^\perp = \mathrm{R}(D_1)^\perp \cap \mathrm{R}(D_2)^\perp \cap \cdots \cap \mathrm{R}(D_e)^\perp, \tag{8}$$

thus, in terms of projections,

$$P = P_1 \wedge P_2 \wedge \cdots \wedge P_e. \tag{9}$$

To see that (8) holds, assume $\eta$ is a vector on the left hand side, i.e. $\eta \in \mathrm{R}(D)^\perp$. Then $\eta$ is orthogonal to the column space of $D$, thus to every column of $D$. Since the columns of $D$ consist of the columns of $D_1, \ldots, D_e$, $\eta$ must be orthogonal to each $\mathrm{R}(D_i)$ for $i = 1..e$; we therefore have $\eta \in \mathrm{R}(D_i)^\perp$, and, in consequence, $\eta$ is in the intersection on the right hand side. Conversely, assume that for each $i = 1..e$, we have $\eta \in \mathrm{R}(D_i)^\perp$, i.e. $\eta$ is in the intersection on the right hand side. For each $i$, $\eta$ is orthogonal to $\mathrm{R}(D_i)$, and thus to every column of $D_i$. Again, since the columns of $D$ consist of columns from the $D_i$s, $\eta$ is orthogonal to every column of $D$, hence $\eta \in \mathrm{R}(D)^\perp$. Ergo, equality holds in (8).

By using (Halperin, 1962, Theorem 1) on (9), we now have

$$P = \lim_{n \to \infty} (P_1 P_2 \cdots P_e)^n. \tag{10}$$

The convergence holds in the strong operator topology, i.e. pointwise on vectors, although for finite dimension, it is equivalent to both the weak and the uniform operator topology.

This shows that the following algorithm converges.

**Algorithm 3.1** (Method of Alternating Projections)**.** *Let $v$ be a vector (typically a column of $X$ or $y$). The following algorithm converges to $Pv$. It is a direct generalization of the within-groups transformation (i.e. the case $e = 1$).*

*(1) Pick a tolerance, an $\epsilon > 0$, e.g. $\epsilon = 10^{-9}$. Let $v_0 = v$, and $i = 0$.*

*(2) Let $z_0 = v_i$. For $j = 1..e$, form $z_j$ by subtracting the group means of the groups in $D_j$ from $z_{j-1}$. I.e. $z_j = P_j z_{j-1}$.*

*(3) Let $v_{i+1} = z_e$. If $\|v_{i+1} - v_i\| < \epsilon$, terminate with the vector $v_{i+1}$ as an approximation to $Pv$. Otherwise, increase $i$ by 1. Go to step (2).*

**Remark 3.2.** By using Algorithm 3.1 on $y$ and the columns of $X$, we find $Py$ and $PX$, and the estimation of $\hat{\beta}$ from (6) is therefore manageable.

**Example 3.3.** *In the case with $e = 2$, i.e. with two fixed effects like "firm" and "individual", Algorithm 3.1 amounts to repeating the process of centring on the firm means, followed by centring on the individual means, until the vector no longer changes.*

### *3.1. Convergence rate*

The rate of convergence for the method of alternating projections is analyzed in Deutsch and Hundal (1997), though by general concepts for which the author has not found an intuitive description in terms of the covariates used to construct $D$ in a typical panel data model. For the case $e = 2$, (Aronszajn, 1950), cited in (Deutsch and Hundal, 1997, Corollary 2.9), has an estimate

$$\|(P_1 P_2)^n - P\| \leq \cos^{2n-1}(\mathrm{R}(D_1)^\perp, \mathrm{R}(D_2)^\perp), \tag{11}$$

where the function cos denotes the cosine of the (complementary) angle between subspaces. The inequality (11) was later shown in (Kayalar and Weinert, 1988,

Theorem 2) to be an equality. The quantity on the right hand side is strictly smaller than 1 in finite dimensional spaces (Deutsch and Hundal, 1997, Lemma 2.3(3)). Thus, we have linear convergence, but the rate varies with the structure of $D$. In (Deutsch and Hundal, 1997, Theorem 2.7 and Section 3), the case $e > 2$ is also handled, but the convergence rate is more complicated.

## 4. Finding the fixed effects

Having found $\hat{\beta}$ as per Remark 3.2, we now look at the problem of finding $\hat{\alpha}$. Recall from (4) that

$$\mathcal{D}'\mathcal{D}\hat{\alpha}_r = \mathcal{D}'\rho, \tag{12}$$

where $\rho = y - X\hat{\beta}$ are the residuals for the original system with dummies omitted.

In the special case with a single category variable ($e = 1$), the columns of $\mathcal{D}$ are orthogonal, and $\mathcal{D}'\mathcal{D}$ is therefore diagonal; $\hat{\alpha}_r$ is simply the group means of the residuals $\rho$. This is the within-groups estimator. If $e > 1$, $\mathcal{D}'\mathcal{D}$ is not diagonal, though it is typically quite sparse, and the pedestrian approach is to apply a good sparse solver on (12), such as those evaluated in Gould et al. (2007). However, there is another avenue which has proven useful.

### *4.1. The Kaczmarz method*

We rewrite (7) as

$$\mathcal{D}\hat{\alpha}_r = (I - P)(y - X\hat{\beta}) = (y - X\hat{\beta}) - (Py - PX\hat{\beta}) \tag{13}$$

where the right hand side is readily computed when we have $\hat{\beta}$. Now, although the derivation of (7) assumed that linearly dependent columns of $D$ had been removed, we may consider the equation with the original rank-deficient $D$. As noted in Remark 2.2, the right hand side will be the same, and since $D$ and $\mathcal{D}$ have the same range, the equation with $D$ will still have a solution. I.e. we consider the equation

$$D\hat{\alpha} = (y - X\hat{\beta}) - (Py - PX\hat{\beta}). \tag{14}$$

where $\hat{\alpha}$ is only identified up to translation by the null-space of $D$. We resolve the ambiguity by applying an estimable function to the solution.

To find a solution to (14), we can apply the Kaczmarz method, (Kaczmarz, 1937). Following the method, we can view each equation in system (14) as defining a hyperplane in $\mathbb{R}^g$. The intersection of all the hyperplanes is the solution set of the system. The Kaczmarz method is a variant of the method of alternating projections, where we start with a vector and successively project it onto each hyperplane. The process is repeated until the vector stops changing. This will have brought us to the intersection, i.e. a solution to (14).

In our case, the projection onto a hyperplane is very simple. Remember that each row of $D$ contains exactly $e$ 1's; the other entries are zero. We write row $i$ of system (14), with $x$ instead of $\hat{\alpha}$, as $\langle d_i, x\rangle = b_i$, where $d_i$ is row $i$ of $D$, $b_i$ is the $i$'th coordinate of the right hand side, and $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. The projection onto the solution set of row $i$ is

$$x \mapsto x - (\langle d_i, x\rangle - b_i)d_i/\|d_i\|^2, \tag{15}$$

which is easy to compute; $\|d_i\|^2 = e$ for every $i$, and the inner product is merely a sum of $e$ of the coordinates of $x$. In other words, the update to $x$ requires minimal computation.

Solving (14) consists of starting with e.g. $x = 0$, and applying the projections (15), for $i = 1..n$ in succession, and repeating this process until the change in $x$ is smaller than some tolerance. A process entirely analogous to Algorithm 3.1, with the $e$ $P_i$s replaced by the $n$ projections in (15). It is noted in (Deutsch and Hundal, 1997, Section 4) that their convergence rate results also cover affine sets, not only subspaces.

**Remark 4.1.** It is easily seen that consecutive duplicate rows in $D$ may be ignored, since the projection in (15) is idempotent.

When we have found a solution $\gamma$ of (14), we must apply an estimable function to $\gamma$, to obtain unique, meaningful coefficients. To see that this works, recall that an estimable function is a matrix operator $F$ whose row space is

contained in the row space of $D$. Denote the null-space of a matrix $M$ by $\mathrm{N}(M)$. From matrix theory, the row-space of $M$ is the orthogonal complement $\mathrm{N}(M)^{\perp}$. Thus, we have $\mathrm{N}(F)^{\perp} \subset \mathrm{N}(D)^{\perp}$, or, equivalently, $\mathrm{N}(D) \subset \mathrm{N}(F)$. Hence, if $\gamma_1$ and $\gamma_2$ are two arbitrary solutions of (14), i.e. $D\gamma_1 = D\gamma_2$, we have $D(\gamma_1 - \gamma_2) = 0$, thus $F(\gamma_1 - \gamma_2) = 0$, so $F\gamma_1 = F\gamma_2$. That is, the value of the estimable function does not depend on the particular solution of (14).

## 5. Identification with two fixed effects

Since the method described above is typically used with category variables with many different values, there is a real chance of spurious relations occurring between them, resulting in non-obvious rank-deficiency in $D$, and, it follows, an identification problem for $\hat{\alpha}$.

To be complete, we now recall a known identification result for the case $e = 2$. This is needed to find estimable functions. Abowd et al. (1999) has analyzed this problem in the case with two dummy-groups (firms and individuals). The problem was also treated much earlier in a different setting, by Eccleston and Hedayat (1974), and the references cited therein.

In Abowd et al.'s approach, an undirected bipartite graph $G$ is constructed in which each vertex consists of a firm or an employee. A firm and an employee are adjacent if and only if the employee has worked for the firm. There are no more edges in the graph. I.e. $D'$, with duplicate columns omitted, is the (vertex-edge) incidence matrix of the graph.

They then analyze identifiability in terms of the connected components (or "mobility groups") of the graph $G$, and show that it is sufficient to have a reference dummy in each of the connected components, (see Abowd et al., 2002, Appendix 1). That is, we have the theorem

**Theorem 5.1** (Various authors)**.** *If $e = 2$, the rank deficiency of $D$, hence of $D'D$, equals the number of connected components of the graph $G$*

*Proof.* To put this result in the context of spectral graph theory, we provide the following reference. The matrix $D'$ may be viewed as the incidence ma-

trix of the graph $G$; $D'D$ is then the *signless Laplacian* of $G$. Moreover, the graph is bipartite, with "firms" in one partition, "employees" in the other. By (Cvetković et al., 2007, Corollary 2.2), the multiplicity of eigenvalue 0, i.e. the rank deficiency, is the number of connected components. □

**Remark 5.2.** For general $e$, we get an $e$-partite, $e$-uniform hypergraph. We are not aware of similar general results for such graphs. However, the case $e = 3$ has been analyzed in Godolphin and Godolphin (2001), who offer a procedure for finding the estimable functions in terms of *staircase partitions*; nevertheless it is not as simple as finding graph theoretic connected components. The spectrum of the signless Laplacian is also an active field of research in graph theory, see Cvetković and Simić (2010) for a survey.

## 6. Summary

We summarize the above results in an algorithm.

**Algorithm 6.1.** *To find OLS estimates $\hat{\beta}$ and $\hat{\alpha}$ for model* (1) *we proceed in the following manner.*

*(1) Centre $\bar{y} = Py$ and $\bar{X} = PX$ with Algorithm 3.1*

*(2) Perform an OLS on*

$$\bar{y} = \bar{X}\beta + \epsilon.$$

*The result of the estimation is $\hat{\beta}$. The covariance matrix must be adjusted, taking into account the number of eliminated parameters in the degrees of freedom, as in Remark 2.1.*

*(3) Compute the quantity $B = (y - X\hat{\beta}) - (\bar{y} - \bar{X}\hat{\beta})$. Apply the Kaczmarz method to find a solution $\gamma$ of the equation $D\gamma = B$, as in* (14).

*(4) Apply an estimable function to $\gamma$, suitable to the problem at hand, to get meaningful estimates $\hat{\alpha}$.*

In step (2), the number of eliminated parameters is the rank of $D$. I.e., if the OLS in step (2) reports the covariance matrix $\Sigma$, and $d$ degrees of freedom, the true degrees of freedom are $d - \operatorname{rank} D$, so that $d(d - \operatorname{rank} D)^{-1}\Sigma$ is the right covariance matrix for $\hat{\beta}$. For $e = 2$ we can use Theorem 5.1 to compute the rank-deficiency, and thus the rank of $D$.

There is no easy way to compute the exact degrees of freedom for $e > 2$, although the rank-deficiency of $D$ is at least $e - 1$; if we take this to be the true value, we will be over-estimating the standard errors, which is better than nothing. The sample is typically very large, also compared to the number of dummies, so the relative error in degrees of freedom will typically be small. Alternatively, one may compute the rank-deficiency of $D$ or $D'D$ by picking a small $\epsilon > 0$, do a sparse, pivoted Cholesky decomposition of $D'D + \epsilon I$, and count small pivots, though it is likely to be very time-consuming due to the size of $D'D$.

In step (4), finding an estimable function in the case $e = 2$, is guided by Theorem 5.1. There is a discussion of this in McCaffrey et al. (2010). To be brief, a difference of two "firms", or two "individuals", in the same component is estimable. Likewise, the sum of a "firm" and an "individual". So, if we pick a reference "individual", we may subtract its value from all the "individuals", and add it to all the "firms", in each component. We can also combine this with subtracting the mean of the "firms" from all the "firms", and use the mean as a common intercept, still for each component. Coefficients from different components are not directly comparable. For the case $e > 2$ no intuitive identification results are known, and thus no easy method is currently known for finding an intuitive estimable function which allows coefficients to be interpreted as partial effects. Having said that, the researcher may know of them in particular cases.

**Remark 6.2.** Note that a candidate estimable function may be tested for non-estimability by running the Kaczmarz method with two randomly chosen initial vectors to get two different solutions to (14). If the function evaluates to two different values, it is not estimable. Since the Kaczmarz step computes the

projection of the initial value on the solution space, two different initial values could possibly produce identical solutions even though the function is not estimable. Say we have a candidate $F$ for an estimable function which actually is not estimable. We have an initial vector 0 which yields the solution $\xi_0$ of (14) and another initial vector $\nu$ which yields the solution $\xi$. If $\xi - \xi_0 \in \mathrm{N}(F)$ we will falsely conclude that the function is estimable. Now, $\xi_0$ and $\xi$ are solutions of (14), so that $\xi - \xi_0 \in \mathrm{N}(D)$. Since $F$ is not estimable, we have by definition, $\mathrm{N}(D) \not\subset \mathrm{N}(F)$, whereas we have $\xi - \xi_0 \in \mathrm{N}(D) \cap \mathrm{N}(F)$. That is, $\xi$ lives in a lower-dimensional space than $\mathrm{N}(D)$, i.e. a set of Lebesgue-measure zero in the solution set of (14). Our test will therefore fail on a set of measure zero.

For the case $e = 2$, if standard errors for $\hat{\alpha}$ are needed, one may bootstrap with Algorithm 6.1. For the case $e > 2$, without an estimable function, it would not be meaningful to bootstrap or otherwise estimate standard errors.

**Remark 6.3.** None of the steps in Algorithm 6.1 are particularly memory-consuming. However, while centring in step (1) can be done in place, the centred vectors $\bar{X}$ and $\bar{y}$ require a copy since we need the original $X$ and $y$ in step (3). The Kaczmarz method in step (3) also consists of operations on single vectors; no large matrices are involved.

## Appendix A. Efficiency compared with other methods

The author has made an implementation of Algorithm 6.1 available on "cran" as an R-package `lfe` (Gaure, 2011). It is a straightforward implementation, i.e. we have not tried to adapt any of the available alternating projections acceleration schemes (e.g., Andersen and Kak, 1984; Gearhart and Koshy, 1989; Martínez and Sampaio, 1986; Strohmer and Vershynin, 2009), nor have we performed extensive machine-related optimization.

The method of Strohmer and Vershynin (2009) works by assigning a probability to each projection, proportional to the $\|d_i\|$ of equation (15), and drawing projections from this distribution rather than applying them in sequence. In

our case, all the probabilities will be the same, and the method will reduce to random shuffling of the equations, a point we discuss below. The method of Andersen and Kak (1984) is geared towards image reconstruction with low visual noise, and make use of the fact that medical image features often are convex. It is unclear to the author whether this could be translated to the problem at hand. The method of Gearhart and Koshy (1989) is a line search method which may well be applicable to the problem at hand, both the centring and the Kaczmarz step, but it has not been implemented in `lfe`. The method of Martínez and Sampaio (1986) is a parallel implementation of the Kaczmarz method. It may be applicable to the Kaczmarz step of `lfe`; the centring is already parallel.

The current implementation in `lfe` suffers from somewhat low floating point ops per cycle ratio due to out-of-cache computations. It has this in common with a lot of other software, including Stata. It could perhaps be improved by interleaving the vectors to be centred, but this would incur additional bookkeeping overheads, and the efficiency of such partakings is prone to be dependent on memory and CPU architecture.

McCaffrey et al. (2010) has compared the various Stata programs for estimating the models in question. We have compared the speed of the fastest method there, a Stata program `a2reg` (Ouazad, 2008), based on the preconditioned conjugate gradient approach of Abowd et al. (2002), to the R-package `lfe`.

The endeavour is not a direct comparison of the algorithms, since implementation details, as well as specifics of the computing contraption, also matter. Nor is it a comprehensive test: the rate of convergence may vary with the structure of the data, according to (11), but it does show that the method presented here is comparable to other generally available methods.

Ideally, a more comprehensive test should be made, but the problem with comprehensive tests on real datasets of this size and form is that the datasets are typically sourced from various public registries and therefore subject to data protection regulations, or not publicly available for commercial reasons. The three relevant datasets in McCaffrey et al. (2010) are of this type. The efficiency

of sparse solvers typically depends on the structure of the underlying density pattern/graph which may be very varied. As an example, the comparison in Gould et al. (2007) uses a total of 149 different matrices. The author has not managed to obtain such a sizeable number of datasets. For the same reason, it is also possible that `a2reg` is not the generally fastest method of the ones compared in McCaffrey et al. (2010).

Because the present method splits the system into separate parts for $\hat{\beta}$ and $\hat{\alpha}$, `lfe` is able to provide standard errors for the $\hat{\beta}$ coefficients, whereas `a2reg` is not. Standard errors for $\hat{\alpha}$ must still be obtained by bootstrapping, though. According to McCaffrey et al. (2010), the other Stata programs providing standard errors are considerably slower than `a2reg` for the large dataset there by a factor of $\approx 40$.

*Simulated datasets*

We have created some simple, simulated job-change panel datasets with two normally distributed covariates $x_1$ and $x_2$. Individuals and firms are assigned constant effects at random. Every individual is observed for 15 observation periods, with a probability of 10% per period of entering a new firm. To obtain variation in firm sizes, the firms are assigned a probability of being chosen drawn from a $\chi^2_{10}$-distribution. We then create an outcome variable as a linear combination of the two covariates and the individual and firm effects, with a normally distributed stochastic term $\epsilon$. I.e., we have

$$y_{itj} = 0.5x_{1it} + 0.25x_{2it} + \lambda_i + \gamma_j + \epsilon_{itj},$$

where $y_{itj}$ is the outcome for individual $i$ who is in firm $j$ at time $t$, the $x$s are the above-mentioned covariates, $\lambda_i$ is the individual fixed effect for individual $i$, and $\gamma_j$ is the firm fixed effect for firm $j$. In the model setup, the dummies for the $\lambda_i$s and $\gamma_j$s enter the matrix $D$ of model (1), whereas the $x$-covariates enter the matrix $X$.

To complicate matters further, and to avoid having balanced datasets, i.e. every individual observed in every period, we have simply taken a 70% sample of

all observations, from which we extracted the largest connected firm/employee component. We created different datasets in the following way.

small $e = 2$. We created a "small" dataset of $\approx 300,000$ individuals, 30,000 different firms, and approximately 3.1 million observations.

big $e = 2$. A "big" dataset contains approximately 4 million individuals, 400,000 firms, with approximately 42 million observations, comparable in size to the labour market of a small country.

wide $e = 2$. In addition, we also put together a "wide" dataset, a modification of the "small" one, but with a "year-effect" thrown in, i.e. an effect for each of the 15 observation periods, with accompanying dummies in $X$.

shoe $e = 3$. To test the speed of `lfe` in the case $e = 3$, we have also created a variation of the "small" dataset where the individuals change their shoes with a probability of 10% per observation period. There are 4,000 types of shoes to choose from, each with their own effect on the wage $y$. Now this is admittedly an unrealistic concoction, and there is little reason to believe that shoe effects would even have the same sign for a sewage worker and an accountant. The shoe dummies enter the $D$ matrix in the estimation. This dataset is referred to as "shoe" below. `a2reg` does not support $e = 3$, so a comparison is not possible unless the 4000 shoe dummies are put into the $X$ matrix.

*Real datasets*

We also have some real datasets. One dataset is from Markussen and Røed (2012), studying peer group effects on social insurance dependency. There are approximately 1 million individuals in this set, observed annually for 16 years, resulting in 16.4 million observations all told. This dataset allows for various dummy structures. All of them use a dummy variable for each individual. The variation is what constitutes the other dummy groups. We have chosen the following variants.

mr1 $e = 2$. There is one other dummy group, the interaction between gender, observation year, cohort, education and region. It contains approximately 1 million different values. There are 5 covariates in the $X$-matrix.

mr2 $e = 2$. The dummy group is the interaction between gender, year, cohort and education. It has 20,311 different values and there are 6 covariates in the $X$-matrix.

mr3 $e = 5$. There are four additional dummy groups. Interactions between region and year (1,321 values), between cohort and year(271), gender and age(29), gender and year(16). There is 1 covariate in the $X$-matrix.

mr4 $e = 4$. This is the same as "mr3", but the gender-age interaction has been placed in the $X$-matrix. Thus, we have $e = 4$, and there are 29 covariates, using one of the gender-age dummies as a reference.

mr5 $e = 4$. This is the same as "mr3", but the gender-year interaction has been placed in the $X$-matrix. This gives us $e = 4$, and there are 16 covariates.

bd1 $e = 2$. This is a different in-house dataset, on wage-formation, with no published results yet. There are approximately 2.7 million individual dummies, 290,000 firm dummies, with 25 million observations. There are 18 covariates in the $X$-matrix. We have extracted the largest connected component of this dataset.

bd2 $e = 2$. Same type of dataset as "bd1", but a different set of individuals, and, consequently, a different sparsity pattern. There are approximately 25 million observations, with 2.7 million individuals and $\approx 310,000$ firms. There are 9 covariates in the $X$-matrix, and 17,558 connected components/mobility groups.

*Comparison*

The comparison has been conducted on a computer running CentOS 5.2, a Dell Poweredge M610 equipped with 48 GiB of 1.33 GHz memory, and 2 hexa-core Intel Xeon X5650 cpus at 2.67 GHz, with an L3-cache of 12 MiB each,

part of the "Titan III" high performance computing facility at the University of Oslo. Stata version 11.0 was used with `a2reg`; R-version 2.11.1 was used with `lfe`. R, and `lfe`, was compiled with the GNU compiler suite, version 4.1.2. The parallel version of the AMD provided blas-library `ACML` 4.4.0 was linked into R, though it was only used for solving the projected system (6).

`lfe` can make use of all the cores in the computer, i.e. it can centre the vectors in parallel. There is nothing to gain from using more cores than there are vectors to centre. We have run `lfe` on a single core, on 3 cores, and 8 cores. We left the thread distribution over the cores to the default NUMA policy in CentOS. We did not try `a2reg` with the multicore-enabled Stata/MP version, and do not know whether it would benefit from this.

The estimates of the "big" dataset, from `a2reg` and `lfe`, of $\hat{\beta}$, and of $\hat{\alpha}$ after applying the same estimable function, corresponding to a single individual reference, were almost equal. All estimated effects were between -5 and 13. An elaboration of "almost equal" is necessary. As both methods purport to solve the normal equations, if only to a tolerance, it is meaningful to measure the differences directly.

The Pearson's correlation between individual effects from the two methods, exceeded $1 - 10^{-10}$. This held true for the firm effects as well. The maximum difference ($\ell^\infty$-distance) between the $\hat{\beta}$s was $10^{-7}$. The differences between the $\hat{\alpha}$s were somewhat larger. The $\ell^\infty$-distance for both firm and individual vectors, was 0.005, the mean difference (dimension-normalized $\ell^1$-distance) was $\approx 10^{-6}$, with a standard deviation of $\approx 10^{-5}$. The Euclidean ($\ell^2$) distance between the firm vectors was 0.008; between the individuals, 0.016. Eight of the firm-estimates, and 26 of the individual estimates differed by more than $10^{-3}$. Differences exceeding $10^{-4}$ were found in 152 of the firms, and 390 of the individuals. For most practical purposes, the estimates from `lfe` and `a2reg` may be considered equal.

The timings in table A.1 are wall-clock timings in seconds for the estimation only, excluding reading and writing of files. Irrelevant cells are filled with "-", i.e. `a2reg` with $e > 2$, and `lfe` with more cores than vectors to centre. Variation

in timing over multiple runs on the same dataset is within 10% for all cases. The reported timings are from a typical run.

Table A.1: Comparison of execution times

| Dataset | $n$ | $k$ | $e$ | `a2reg` | `lfe`(1 core) | `lfe`(3) | `lfe`(8) |
|---|---|---|---|---|---|---|---|
| small | 3.1M | 2 | 2 | 151 | 72( 8) | 56( 8) | - |
| wide | 3.1M | 16 | 2 | 383 | 235( 7) | 108( 6) | 78(6) |
| big | 42M | 2 | 2 | 3168 | 1980(140) | 1186(148) | - |
| shoe | 3.1M | 2 | 3 | - | 104( 30) | 90( 29) | - |
| mr1 | 16.4M | 5 | 2 | 1159 | *470*( ??) | *280*( ??) | *228*( ??) |
| mr2 | 16.4M | 6 | 2 | 1435 | 667(216) | 498(215) | 430(217) |
| mr3 | 16.4M | 1 | 5 | - | *49745*( ??) | *31779*( ??) | - |
| mr4 | 16.4M | 29 | 4 | - | 6508(143) | 2434(143) | 1496(145) |
| mr5 | 16.4M | 16 | 4 | - | 12573(884) | 5947(915) | 3993(923) |
| bd1 | 24.7M | 18 | 2 | 3966 | 20660(559) | 9363(552) | 7515(568) |
| bd2 | 25.4M | 9 | 2 | 2760 | 4234(133) | 1833(133) | 1259(132) |

The figures in brackets for `lfe` is the time spent in the Kaczmarz step.

*Discussion of comparison*

The results are mixed. For some datasets, `lfe` outperforms `a2reg`, for other datasets it's the other way around. In particular for the dataset "bd1", and, importantly, the `lfe`-figures for the "mr1" and "mr3" datasets do not include numbers for the Kaczmarz step, since the step *failed to complete in 20 hours*. The figures for these two datasets are for the centring and projected OLS only. This, it should be noted, is an extreme case. The dataset was not constructed for the purpose of estimating parameters for these dummies; they are simply control variables. However, it does illustrate a potential shortfall of the Kaczmarz method. Also, the difference between "mr3", "mr4" and "mr5", which have identical normal equations but with a different one of the dummy sets of "mr3" moved into the $X$-matrix in "mr4" and "mr5", is striking. Similarly, the

difference between the `lfe`-timings for "bd1" and "bd2". Significantly, `a2reg` does not seem to exhibit such a strong structural dependence. This shows that not only in the theory of eq (11), but also in practice, the execution time of both `a2reg` and `lfe` depends, possibly in inherently different ways, on the size *and* the structure of the data.

Now the fact that the parallel speedup for `lfe` is not linear is due to Amdahl's law. That is to say, only the actual centring was parallelized, whereas the projected OLS, the Kaczmarz step, and mundane bookkeeping, such as creating a "model matrix" from the "data frame", ran serially. The actual centring of each of the 3 vectors in the "big" dataset on 1 core took 376, 410, and 414 seconds. On 3 cores, centring took 484 seconds. The projected OLS took, unsurprisingly, only 8 seconds. So, centring, projected OLS, and the Kaczmarz step, add up to 640 seconds, leaving 546 seconds for various bookkeeping jobs.

Since centring is such a simple operation, and the vector length exceeds the cache size, parallel centring over many cores, in the current implementation, is likely to exceed the total memory bandwidth, which also tends to limit parallel efficiency.

Memory usage was higher for `lfe` than for `a2reg`, both because the algorithm requires a copy of $X$ and $y$, as in Remark 6.3, but also because R is a functional programming language with some implicit copying and garbage collection, requiring in effect another temporary copy of $X$ and $y$, and even more. Memory usage after reading the "big" dataset, but before starting computations, was $\approx$ 3 GiB in Stata, and 6 GiB in R. The Stata estimation peaked at approximately 10 GiB total memory, whereas the R estimation peaked at 13 GiB.

*A note on convergence*

We have used the following, intuitively derived, modification of the convergence criterion for the alternating projection methods. Let $x_0$ be the initial vector. Let $x_i = T^i x_0$ where $T$ is the product of projections. The change $\delta_i = \|x_{i+1} - x_i\|$ should be decreasing with $i$. We compute $c_i = \delta_i/\delta_{i-1}$, which

should be less than 1. We then decide to have converged when $\delta_i < \epsilon(1 - c_i)$ for a pre-specified $\epsilon = 10^{-8}$, but convergence to have failed if $c_i \geq 1$. Intuitively, if $c_i = c$ were constant, then the sum of the remaining updates, $\sum_{j=i}^{\infty} \delta_j = \delta_i/(1 - c)$, should be less than $\epsilon$. If the change stops shrinking, we might arguably be running in circles due to numerical inaccuracy. Setting $\epsilon = 0$ will cause iterations to continue until machine precision has been reached.

When estimating the "wide" dataset, we have put the time dummies in the $X$-matrix with a single reference, thus keeping $e = 2$. We also tried to put them in the $D$ matrix, to get $e = 3$, and guessed on an estimable function, equivalent to what we obtain when removing a single time reference. This approach leads to convergence problems in the Kaczmarz step. Although it produced the same $\hat{\beta}$ results (for the $x_1$, $x_2$), and year-effects within an $\ell^\infty$-distance of $10^{-3}$ of the $e = 2$ case, the firm and individual effects were poorer, with $\ell^\infty$-distance of 1.15, and correlation with the $e = 2$ case of a mere 0.985.

Having the time dummies in $D$, although it reduces the number of vectors to be centred from 17 to 3, slows the convergence (per vector) of the centring somewhat, and the Kaczmarz step dramatically, from 7 to 1,099 seconds, ending in convergence failure after $\approx 30,000$ iterations. One obvious reason for the tardiness is the absence of repeated rows in the enlarged $D$, as in Remark 4.1. This could only explain a slow-down factor of up to 15, the number of observations for each individual, so it is not the full explanation. However, system (14) has also become more redundant. Since the right hand side is numerically determined to a tolerance only, and each projection in the Kaczmarz step is also done with finite precision, there is a real risk that the system may become numerically inconsistent, leading the Kaczmarz method astray.

However, it turns out that the order of the equations has a significant impact on the convergence properties. In the test runs above, the dataset, and thus the rows in (14), was ordered by "individual" and "year". By ordering the rows in (14) randomly prior to the Kaczmarz iterations, the "wide" dataset with the time dummies in $D$, finished in 152 seconds on 1 core, and 17 seconds in the Kaczmarz step. This is a simple variant of the scheme described by Strohmer

and Vershynin (2009), and was noted in Natterer (1986).

A similar effect is present in the "shoe" dataset; ordering the observations randomly, the time spent in the Kaczmarz step fell from 30 to 15 seconds.

We have not analyzed this phenomenon in any greater detail, and we do not know if random shuffling would impede the Kaczmarz convergence. It did not help with our "mr1" dataset. But our recommendation in the event of convergence problems, is to order the equations in the Kaczmarz method randomly.

We should also keep in mind the words of Strohmer and Vershynin (2009):

> *Despite the popularity of this method, useful theoretical estimates for its rate of convergence are still scarce.*

**Acknowledgements**

I wish to thank Bjørn-Helge Mevik, the editors, and five anonymous reviewers for valuable comments. I also wish to thank Simen Markussen and Bjorn Dapi for providing datasets for development and testing.

**References**

Aaronson, D., Barrow, L., 2007. Teachers and Student Achievement in the Chicago Public High Schools. Journal of Labor Economics 25, 95–135.

Abowd, J.M., Creecy, R.H., Kramarz, F., 2002. Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data. Technical Report TP-2002-06. U.S. Census Bureau.

Abowd, J.M., Kramarz, F., Margolis, D.N., 1999. High Wage Workers and High Wage Firms. Econometrica 67, 251–333.

Abowd, J.M., Kramarz, F., Roux, S., 2006. Wages, Mobility and Firm Performance: Advantages and Insights from Using Matched WorkerFirm Data. The Economic Journal 116, 245–285.

Andersen, A., Kak, A., 1984. Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm. Ultrasonic Imaging 6, 81–94.

Andrews, M., Gill, L., Schank, T., Upward, R., 2008. High wage workers and low wage firms: negative assortative matching or limited mobility bias? J.R. Stat. Soc.(A) 171(3), 673–697.

Ansley, C.F., Kohn, R., 1994. Convergence of the Backfitting Algorithm for Additive Models. J Austral Math Soc 57, 316–329.

Aronszajn, L., 1950. Theory of Reproducing Kernels. Trans. Amer. Math. Soc. 68, 337–404.

Bazzoli, G.J., Chen, H.F., Zhao, M., Lindrooth, R.C., 2008. Hospital financial condition and the quality of patient care. Health Economics 17, 977–995.

Carneiro, A., Guimaraes, P., Portugal, P., 2012. Real Wages and the Business Cycle: Accounting for Worker and Firm Heterogeneity. American Economic Journal: Macroeconomics 4, 133–152.

Cornelißen, T., 2008. The stata command felsdvreg to fit a linear model with two high-dimensional fixed effects. Stata Journal 8, 170–189(20).

Cornelißen, T., Hübler, O., 2011. Unobserved individual and firm heterogeneity in wage and job-duration functions: Evidence from german linked employeremployee data. German Economic Review 12, 469–489.

Cvetković, D., Rowlinson, P., Simić, S.K., 2007. Signless Laplacians of finite graphs. Linear Algebra and its applications 423, 155–171.

Cvetković, D., Simić, S.K., 2010. Towards a spectral theory of graphs based on the signless Laplacian, III. Appl. Anal. Discrete Math. 4, 156–166.

Deutsch, F., Hundal, H., 1997. The Rate of Convergence for the Method of Alternating Projections, II. J. Math. Anal. App. 205, 381–405.

Eccleston, J.A., Hedayat, A., 1974. On the Theory of Connected Designs: Characterization and Optimality. Ann. Statist. 2, 1238–1255.

Gaure, S., 2011. lfe: Linear Group Fixed Effects. R package version 1.4-641.

Gearhart, W.B., Koshy, M., 1989. Acceleration schemes for the method of alternating projections. Journal of Computational and Applied Mathematics 26, 235–249.

Gibbons, S., Overman, H.G., Pelkonen, P., 2010. Wage Disparities in Britain: People or Place? SERC Discussion Papers 0060. Spatial Economics Research Centre, LSE.

Godolphin, J.D., Godolphin, E.J., 2001. On the connectivity of row-column designs. Util. Math. 60, 51–65.

Gordon, R., Bender, R., Herman, G.T., 1970. Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography. Journal of Theoretical Biology 29, 471–481.

Gould, N.I.M., Scott, J.A., Hu, Y., 2007. A numerical evaluation of sparse direct solvers for the solution of large sparse symmetric linear systems of equations. ACM Trans. Math. Softw. 33.

Halperin, I., 1962. The Product of Projection Operators. Acta Sci. Math. (Szeged) 23, 96–99.

Herman, G.T., 2009. Fundamentals of computerized tomography: Image reconstruction from projections. Springer. 2 edition.

Hudson, H., Larkin, R., 1994. Accelerated image reconstruction using ordered subsets of projection data. Medical Imaging, IEEE Transactions on 13, 601–609.

Jacob, B.A., Lefgren, L., 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. Journal of Labor Economics 26, pp. 101–136.

Kaczmarz, A., 1937. Angenäherte Auflösung von Systemen linearer Gleichungen. Bulletin International de l'Academie Polonaise des Sciences et des Lettres 35, 355–357.

Kayalar, S., Weinert, H., 1988. Error Bounds for the Method of Alternating Projections. Math. Control Signals Systems 1, 43–59.

Lee, S., Huang, J.Z., 2013. A coordinate descent mm algorithm for fast computation of sparse logistic pca. Computational Statistics & Data Analysis 62, 26–38.

Margolis, D.N., Salvanes, K.G., 2001. Do Firms Really Share Rents with Their Workers? IZA Discussion Paper 330. Institute for the Study of Labor (IZA).

Markussen, S., Røed, K., 2012. Social Insurance Networks. IZA Discussion Papers 6446. Institute for the Study of Labor (IZA).

Martínez, J., Sampaio, R.J.B.D., 1986. Parallel and sequential Kaczmarz methods for solving underdetermined nonlinear equations. Journal of Computational and Applied Mathematics 15, 311–321.

McCaffrey, D.F., Lockwood, J., Mihaly, K., Sass, T.R., 2010. A Review of Stata Routines for Fixed Effects Estimation in Normal Linear Models. Mimeo.

Natterer, F., 1986. The Mathematics of Computerized Tomography. Wiley, New York.

von Neumann, J., 1949. On Rings of Operators. Reduction Theory. Ann. Math. 50, 401–485.

Ouazad, A., 2008. A2REG: Stata module to estimate models with two fixed effects. Statistical Software Components, Boston College Department of Economics.

Schmieder, J.F., 2009. GPREG: Stata module to estimate regressions with two dimensional fixed effects. Statistical Software Components, Boston College Department of Economics.

Strohmer, T., Vershynin, R., 2009. A Randomized Kaczmarz Algorithm with Exponential Convergence. Journal of Fourier Analysis and Applications 15, 262–278.

Ueki, M., Kawasaki, Y., 2013. Multiple choice from competing regression models under multicollinearity based on standardized update. Computational Statistics & Data Analysis 63, 31–41.

Vidaurre, D., Bielza, C., Larrañaga, P., 2013. Sparse regularized local regression. Computational Statistics & Data Analysis 62, 122–135.

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. The MIT Press.