Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis

Anne Bernard, *CNAM, CEDRIC Paris, CE.R.I.E.S.*, anne.bernard@ceries-lab.com
Christiane Guinot, *CE.R.I.E.S.*, christiane.guinot@ceries-lab.com
Gilbert Saporta, *CNAM, CEDRIC Paris*, gilbert.saporta@cnam.fr

Abstract. Two new methods to select groups of variables have been developed for multiblock data: "Group Sparse Principal Component Analysis" (GSPCA) for continuous variables and "Sparse Multiple Correspondence Analysis" (SMCA) for categorical variables. GSPCA is a compromise between Sparse PCA method of Zou, Hastie and Tibshirani and the method "group Lasso" of Yuan and Lin. PCA is formulated as a regression-type optimization problem and uses the constraints of the group Lasso on regression coefficients to produce modified principal components with sparse loadings. It leads to reduce the number of nonzero coefficients, i.e. the number of selected groups. SMCA is a straightforward extension of GSPCA to groups of indicator variables, with the chi-square metric. Two real examples will be used to illustrate each method. The first one is a data set on 25 trace elements measured in three tissues of 48 crabs (25 blocks of 3 variables). The second one is a data set of 502 women aimed at the identification of genes affecting skin aging with more than 370.000 blocks, each block corresponding to SNPs (Single Nucleotide Polymorphisms) coded into 3 categories.

Keywords. Sparse principal component analysis, group Lasso, group variable selection, dimension reduction.

1 Introduction

Variable selection and dimensionality reduction are necessary in different application domains and more specifically in genetics (high dimensional data) to reduce the number of variables and obtain a better interpretation of the results. When data are structured by blocks of variables, multiblock data analysis are performed. But in case of high dimensional data with thousands of blocks, the links between blocks (in unsupervised cases) or between a dependent variable and blocks of explanatory variables (in supervised case) is difficult to interpret. A solution to this problem is to select groups of variables to reduce the number of explanatory blocks and find relevant variables. In this paper, we will only consider the unsupervised case. Thus, it is important to find a compromise between the selection of individual input variables (Sparse Principal Component Analysis method) and selection of grouped variables (method "group Lasso").

The Sparse Principal Component Analysis (SPCA) introduced by Zou, Hastie and Tibshirani [1] is a method used to reduce the number of continuous variables in unsupervised cases. Each principal component (PC) of a PCA is a linear combination of the p variables and loadings can be recovered by regressing the PC on the p variables. Therefore, SPCA can consider PCA as a regression-type optimization problem integrating the elastic net constraint into the regression criterion that leads to some zero loadings (sparse loadings). Let the data matrix X be a $n \times p$ matrix, where n and p are the number of observations and the number of variables, respectively. If PCA is computed via the singular value decomposition, $X = UDV^T$ with $Z = UD$ the PCs and V the matrix of the corresponding loadings of the PCs. The iterative SPCA algorithm consists in minimizing the following criterion:

$$\beta^{k*} = \underset{\beta^k}{\operatorname{argmin}} \|Y^k - X\beta^k\|^2 + \lambda\|\beta^k\|^2 + \lambda_{1,k}\|\beta^k\|_1$$

with $Y^k = X\alpha^k$ when $\alpha^k = V[,k]$ the loadings of the k principal components. After applying the algorithm, some component weights are set to zero which reduces the number of explanatory variables and make it easier to interpret the derived PCs.

The "group Lasso" method introduced by Yuan and Lin [2] is an extension of the Lasso for factor selection. It considers the general regression problem with a penalty function which is an intermediate between the l_1 penalty used in Lasso and l_2 penalty used in ridge regression to select groups of variables. Let Y be a $n \times 1$ response variable, X_j a $n \times p_j$ matrix corresponding to the predictors of the j^{th} group, and β_j a coefficient vector of size p_j, $j = 1, ..., J$. Otherwise, the penalty function introduced previously is defined as follows: for a vector $\beta \in R^d$ and H a symmetric d by d positive definite matrix:

$$\|\beta\|_H = (\beta' H \beta)^{1/2}.$$

Thus, the group Lasso estimate is defined as the solution to:

$$\tfrac{1}{2}\|Y - \sum_{j=1}^{J} X_j\beta_j\|^2 + \lambda \sum_{j=1}^{J} \|\beta_j\|_{H_j}$$

where $\lambda \geq 0$ is a tuning parameter and $H_1, ..., H_J$ positive definite matrices. As proposed by Yuan and Lin, we will set $H_j = p_j I_{p_j}$ (I_{p_j} the identity matrix $p_j \times p_j$).

The sections are organized as follows. Group Sparse PCA criterion for continuous variables is defined in section 2 and its generalization (Sparse Multiple Correspondence Analysis) for categorical variables is introduced in section 3. Two real data sets are given in example in section 4 to illustrate both methods: one with continuous variables and another one with categorical variables. Finally, a summary and discussion are given in section 5.

2 Group sparse PCA

Let X be a $n \times p$ data matrix, where n and p are the number of observations and the number of variables, respectively. Suppose that the p predictors are divided into J groups, with p_j the number of predictors in group j, $j = 1, ..., J$. $X = (X_1, ..., X_J)$ with X_j a $n \times p_j$ matrix corresponding to the predictors of the j^{th} group. Assume that X is centered so the observed mean is 0. Group Sparse Principal Component Analysis (GSPCA) method is a compromised between SPCA method and group Lasso. We want to select groups of continuous variables in unsupervised cases. The elastic net penalty function introduced in the SPCA algorithm is replaced by the penalty function defined in the group Lasso to set all weights of an entire block to zero (selection of grouped variable instead of individual variables). Other types of possible penalities are discussed in [3].

Group sparse PCA algorithm

1. Let α start at $V[, 1 : K]$, the loadings of the first K ordinary principal components.
2. Given a fixed α=$[\alpha^1, ..., \alpha^K]$, solve the group lasso problem for $k = 1, ..., K$ (number of factors, $K \leq J$) and j=1$, ..., J$ (number of groups)

$$\beta_j^k = \underset{\beta}{\text{argmin}} \tfrac{1}{2} \|Y^k - \sum_{j=1}^{J} X_j \beta_j^k\|^2 + \lambda \sum_{j=1}^{J} \|\beta_j^k\|_{H_j},$$

   with $Y^k = X\alpha^k$ and λ the tuning parameter (note that the sparsity of the solution is determined by the magnitude of this chosen tuning parameter).
   There are many possible choices for matrices H_j but here we will choose $H_j = p_j I_{p_j}$.
3. For a fixed $B = [\beta^1, ..., \beta^K]$ with $\beta^k = (\beta_1^{\prime k}, ..., \beta_J^{\prime k})$, $k = 1, ..., K$, compute the SVD of $X^T X B = UDV^T$ and update $\alpha = UV^T$.
4. Repeat step 2-3 until convergence.
5. Normalization : $V_j^* = \frac{\beta_j}{\|\beta_j\|}$, $j = 1, ..., J$.

To solve the problem in step 2, the algorithm is based on Nesterov's method for generalized gradient descent and the optimal solution is characterized by the subgradient equations (see [4]).

3 Sparse MCA

Sparse Multiple Correspondence Analysis (SMCA) is a straightforward extension of GSPCA for groups of indicator variables, with the chi-square metric. The SMCA algorithm is based on the same main idea as the GSPCA algorithm, however we don't consider a PCA of the original matrix but a weighted PCA of the matrix of the column and row profiles with the chi-square metric (MCA, [5]) . Let X be a the $n \times J$ matrix of qualitative variables with p_j the number of modalities of the j^{th} variable, $j = 1, ..., J$. MCA begins by constructing the $n \times q$ matrix K of indicator variables (called complete disjunctive table) with $K = (K_1, ..., K_J)$. Each group K_j is composed of the indicator variables of all categories of each qualitative variable. The $n \times q$

matrix of the frequencies F is deduced (f_{is}=$\frac{1}{nJ}$ if i has the modality s, 0 otherwise).
Let $r \in \mathbb{R}^n$ and $c \in \mathbb{R}^q$ be the vectors of the marginal sums of the rows and the columns, respectively and $D_r = diag(r)$ and $D_c = diag(c)$. Therefore, $R = D_r^{-1}(F - rc^t)D_c^{-1}$ becomes the starting point of the SVD. It can be presented as well: $R = (R_1, ..., R_J)$ with R_j a $n \times p_j$ matrix corresponding to the j^{th} variable. The SMCA algorithm take into account the R matrix unlike the GSPCA algorithm that use the original matrix X. The contribution of selected groups for the construction of the axes can be computed and the most important variables among those selected by the SMCA algorithm can be highlighted.

Sparse MCA algorithm

Consider u,v and d be the result of the SVD of $D_r^{1/2} R D_c^{1/2}$. Then set $U = D_r^{-1/2}u$ and V=$D_c^{-1/2}v$.

1. Let α start at $V[, 1 : K]$, for the first K axes.

2. Given a fixed α=$[\alpha^1, ..., \alpha^K]$, solve the group lasso problem for $k = 1, ..., K$ (number of axes, $K \leq J$) and j=$1, ..., J$ (number of groups)

$$\beta_j^k = \underset{\beta}{\text{argmin}} \tfrac{1}{2} \|Y^k - \sum_{j=1}^{J} R_j \beta_j^k\|^2 + \lambda \sum_{j=1}^{J} \|\beta_j^k\|_{H_j}$$

   with $Y^k = R\alpha^k$ and λ the tuning parameter.
   There are many possible choices for matrices H_j but here we will choose $H_j = p_j I_{p_j}$.

3. For a fixed $B = [\beta^1, ..., \beta^K]$ with $\beta^k = (\beta_1'^k, ..., \beta_J'^k)$, $k = 1, ..., K$, compute the SVD of $R^T RB = UDV^T$ and update $\alpha = UV^T$.

4. Repeat step 2-3 until convergence.

5. Normalization : $V_j^* = \frac{\beta_j}{\|\beta_j\|}$, $j = 1, ..., J$.

4 Examples

Continuous variables: Blue Crabs data

The dataset on blue crabs (Callinectes sapidus) has been introduced by Gemperline [6] and analyzed by Kroonenberg [7]. To investigate whether trace element levels were associated with the occurrence of an infection (chitinoclastic bacteria) in blue crabs, tissues of gill, hepatopancreas and muscle were sampled from 48 crabs. Twenty-five trace elements were analysed: $n = 48$ observations (crabs) and $p = 75$ predictors (25 blocks of trace elements from 3 tissues). We will consider the first two principal components (PCs). Figure 1 shows the number of zero loadings on the first two PCs depending on λ: the higher the value of λ, the lower the number of selected blocks. Figure 2 illustrates the selection of trace elements for different values of λ. For a given value of λ, a trace element is selected when the horizontal line coming from this λ value cross through the dots representing this trace element: full dots for the first component and empty

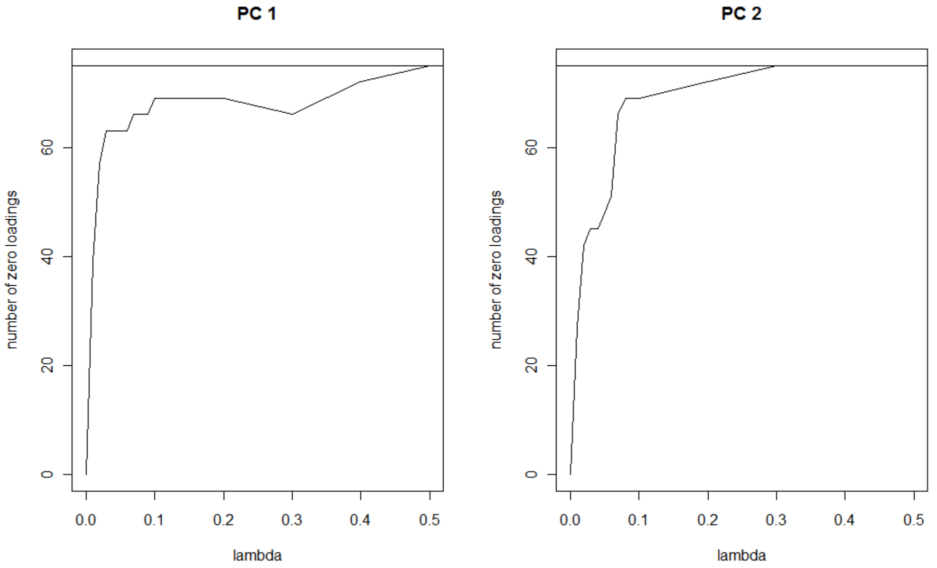


Figure 1: Number of zero loadings computed by GSPCA on the first two PCs depending on λ

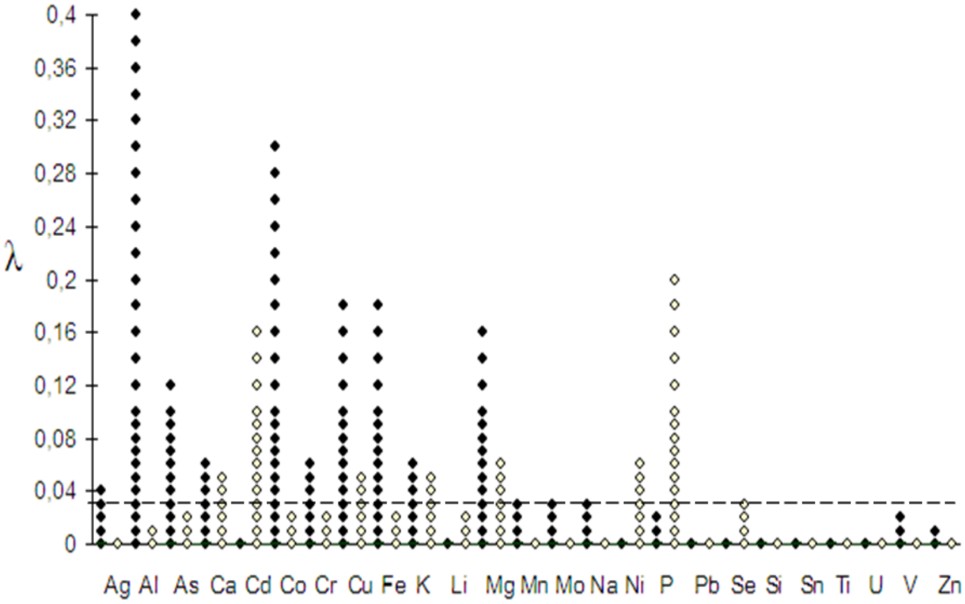
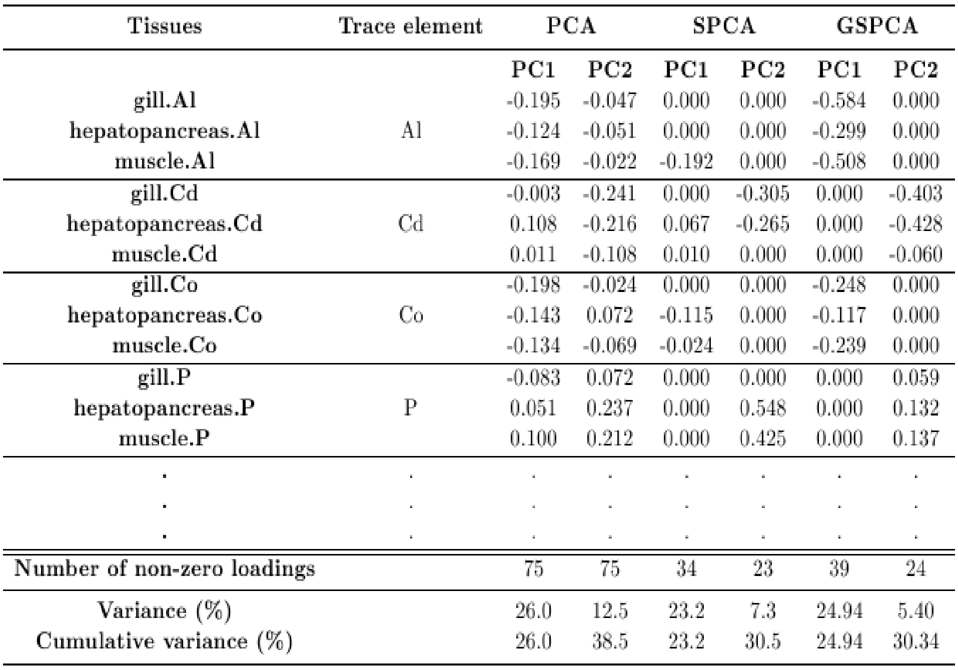


Figure 2: Selected trace elements (blocks) by GSPCA on PC1 (full dots) and PC2 (empty dots) depending on the values of λ



Tissues	Trace element	PCA		SPCA		GSPCA	
		PC1	PC2	PC1	PC2	PC1	PC2
gill.Al	Al	-0.195	-0.047	0.000	0.000	-0.584	0.000
hepatopancreas.Al		-0.124	-0.051	0.000	0.000	-0.299	0.000
muscle.Al		-0.169	-0.022	-0.192	0.000	-0.508	0.000
gill.Cd	Cd	-0.003	-0.241	0.000	-0.305	0.000	-0.403
hepatopancreas.Cd		0.108	-0.216	0.067	-0.265	0.000	-0.428
muscle.Cd		0.011	-0.108	0.010	0.000	0.000	-0.060
gill.Co	Co	-0.198	-0.024	0.000	0.000	-0.248	0.000
hepatopancreas.Co		-0.143	0.072	-0.115	0.000	-0.117	0.000
muscle.Co		-0.134	-0.069	-0.024	0.000	-0.239	0.000
gill.P	P	-0.083	0.072	0.000	0.000	0.000	0.059
hepatopancreas.P		0.051	0.237	0.000	0.548	0.000	0.132
muscle.P		0.100	0.212	0.000	0.425	0.000	0.137
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Number of non-zero loadings		75	75	34	23	39	24
Variance (%)		26.0	12.5	23.2	7.3	24.94	5.40
Cumulative variance (%)		26.0	38.5	23.2	30.5	24.94	30.34

Table 1: Results of the blue crabs data: loadings and variance

dots for the second. For example, for $\lambda = 0.03$, the selected blocks are those whose corresponding dots are located on the horizontal dotted line (i.e., 13 blocks on the first component and 8 blocks on the second). The two last trace elements selected on the first component with a high value of λ are Aluminium (Al) and Cobalt (Co), and on the second Cadmium (Cd) and Phosphor (P). Table 1 summarizes the PCs loadings and the modified PCs loadings computed by SPCA and GSPCA for these 4 groups of variables. They are obtained using tuning parameters adapted for each method. For SPCA, we set $\lambda = 0$ and $\lambda_1 = (0.01, 0.2)$ such that each sparse approximation explains the same variance as PC does in the classical PCA (26.0% vs 23.2% for the first PC), so we set $\lambda = 0.03$ for GSPCA (26.0% vs 24.9%). SPCA correctly identify the most representative variables found with PCA on the two first components and GSPCA selects the entire corresponding groups of variables and puts the others to 0. The adjusted variance is nearly the same for the three methods, but GSPCA produces a much sparser loading structure that makes the interpretation easier.

Categorical variables: SNPs data

We illustrate the SMCA method on a dataset of SNP's (Single Nucleotide Polymorphisms). These data come from a study that has been conducted on 502 women to identify genes affecting skin aging [8]. A blood sample was taken for genetic analysis purposes. The extracted DNA was analyzed using a chip Illumina Human Omni1-Quad containing 1.140.000 genetic markers. More than 370.000 SNPs genotyped have been found in more than 15.198 genes. We will focus on 640 SNPs found in 13 genes previously studied in a candidate gene approach. Each SNP has two alleles and 2 or 3 modalities. We set X a $n \times J$ matrix of categorical variables (SNPs) with

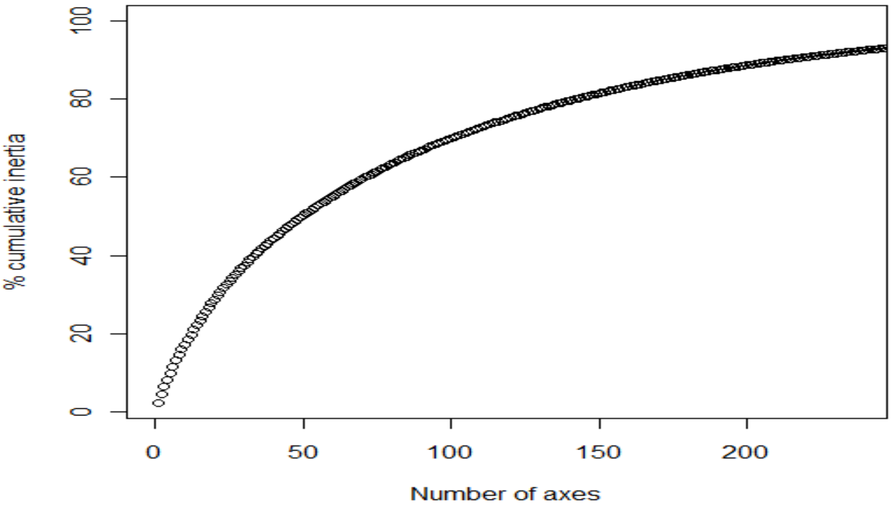


Figure 3: Percentage of cumulative inertia computed by a MCA

$n = 502$ and $J = 640$, and p_j the number of modalities of the j^{th} variable, $j = 1, ..., 640$. We constructed K, a $n \times q$ matrix of the dummy indicator variables for each category of each variable $(q = p_1 + ... + p_J)$. So, K is a 502×1.857 complete disjunctive matrix, with $K = (K_1, ..., K_{640})$ and K_j a $n \times p_j$ matrix. A Multiple Correspondence Analysis (MCA) has been realized on X. The percentage of cumulative inertia is presented in figure 2. We will focus on the first four axes that explained 2.31 %, 2.03%, 1.99% and 1.77% of the total inertia, respectively. As the interpretation of the axes with 1857 variables is not possible, the SMCA method has been applied on the SNPs data to select among the groups of variables (SNPs) the most promising ones. Genetic data are very hard to analyze because of the weakness of the signal coming out of SNPs. It is why the reduction dimension and the variable selection are essential in this application domain. For a low value of λ (0.01), more than 500 SNPs among the 640 available SNPs have been selected by the SMCA without loss of percentage of inertia. A comparison of the results with others values of λ is currently carried out. This step of SNPs selection is very useful in the establishment of a multiblock model in a supervised case to study causal links between skin aging and genetic polymorphisms. The analysis will be subsequently extended to the 370.000 SNPs of the database.

5 Discussion

The GSPCA method has been developed for continuous variables, and SMCA for categorical variables in a unsupervised multiblock data context. Both methods produce sparse loading structures (with limited loss of explained variance) that make easier the interpretation and the comprehension of models. Both are also very powerful in a context of variable selection in high dimension issues. The first example given in Section 4, illustrated the GSPCA method on a small data set, but these methods become meaningful in cases of large data sets as they limit noise as well as computation time. However, these two new methods do not yield sparsity within a group. Therefore, an extension of GSPCA and SMCA could be done in order to select groups and predictors within a group, and so, to produce sparsity at both the group and individual feature levels. Indeed, the selection of one modality of a SNP could be more relevant than the selection of all the modalities of a SNP. This extension would be a compromise between the GSPCA developed here and the new method ”sparse group lasso” developed by Simon et al. [4]. We are expecting that such extension will enable a better and more accurate interpretation of the results, especially in the field of genetic.

Bibliography

[1] Zou, H., Hastie, T. and Tibshirani, R. (2004) *Sparse Principal Component Analysis.* Journal of Computational and Graphical Statistics, **15**, 265–286.

[2] Yuan, M., and Lin, Y. (2006) *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society, Series B, **68**, 49–67.

[3] Van Deun, K. et al. (2011) *A flexible framework for sparse simultaneous component based data integration.* BMC Bioinformatics, **12**, 448.

[4] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2012) *A Sparse-Group Lasso.* Journal of Computational and Graphical Statistics, DOI:10.1080/10618600.2012.681250, Available online: 15 May 2012.

[5] Greenacre, M. (1984) *Theory and applications of correspondence analysis.* London: Acadamic Press.

[6] Gemperline, P. J. et al. (1992) *Principal component analysis, trace elements, and blue crab shell disease.* Analytical Chemistry, **64**, 523–531.

[7] Kroonenberg, P. M. (2008) *Applied multiway data analysis.* Hoboken NJ: Wiley.

[8] Hercberg, S. et al. (2004) *The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals.* Archives Internal Medicine, **164**, 2335–2342.