Appendix to

"Effect fusion using model-based clustering"

Gertraud Malsiner-Walli [2], Daniela Pauger [1], and Helga Wagner [1]

[1] Department of Applied Statistics, Johannes Kepler University Linz, Austria

[2] Institute for Statistics and Mathematics, Vienna University for Economics and Business, Austria

---

Address for correspondence: Gertraud Malsiner-Walli, Institute for Statistics and Mathematics, WU Vienna University for Economics and Business, Welthandelsplatz 1, AT–1020 Wien, Austria.

E-mail: `gmalsine@wu.ac.at`.

Phone: (+43) 1 31336 5595.

Fax: –.

Abstract: This Appendix provides additional material to the paper "Effect fusion using model-based clustering" by Gertraud Malsiner-Walli, Daniela Pauger, and Helga Wagner.

Key words: categorical covariate; sparse finite mixture prior; sparsity; MCMC sampling.

A MCMC sampling

Let $\boldsymbol{b}_0(\boldsymbol{S})$ and $\boldsymbol{B}_0(\boldsymbol{S})$ denote the mean vector and the covariance matrix of the vector of all regression effects β_{jk} conditional on their component indicators S_{jk}, i.e.

$$\boldsymbol{\beta}|\boldsymbol{S} \sim \mathcal{N}(\boldsymbol{b}_0(\boldsymbol{S}), \boldsymbol{B}_0(\boldsymbol{S})),$$

where $\boldsymbol{b}_0(\boldsymbol{S}) = (0, \mu_{1S_{11}}, \dots, \mu_{1S_{1L_1}}, \dots, \mu_{JS_{J1}}, \dots, \mu_{JS_{JL_J}})$ and $\boldsymbol{B}_0(\boldsymbol{S})$ is a diagonal matrix with entries $(\psi_0, \psi_{1S_{11}}, \dots, \psi_{1S_{1L_1}}, \dots, \psi_{JS_{J1}}, \dots, \psi_{JS_{JL_J}})$. Posterior inference using MCMC sampling iterates the following steps:

Regression steps

1. Sample the regression coefficients $\boldsymbol{\beta}$ conditional on $\boldsymbol{S}$ from the normal posterior $\mathcal{N}(\boldsymbol{b}_N, \boldsymbol{B}_N)$, where

$$\boldsymbol{B}_N = \sigma^2(\mathbf{X}'\mathbf{X} + \sigma^2 \boldsymbol{B}_0(\boldsymbol{S})^{-1})^{-1}$$
$$\boldsymbol{b}_N = \boldsymbol{B}_N(\mathbf{X}'\mathbf{y}/\sigma^2 + \boldsymbol{B}_0(\boldsymbol{S})^{-1}\boldsymbol{b}_0(\boldsymbol{S})).$$

2. Sample the error variance σ^2 from its full conditional posterior distribution $\mathcal{G}^{-1}(s_N, S_N)$, where

$$s_N = s_0 + N/2$$
$$S_N = S_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Model based clustering steps

4. For $j = 1, ..., J$ sample the component weights $\boldsymbol{\eta}_j$ from the Dirichlet distribution $Dir(e_{j0}, e_{j1}, \dots, e_{jL_j})$, where

$$e_{jl} = e_0 + N_{jl}, \quad l = 0, \dots, L,$$

and N_{jl} is the number of regression coefficients β_{jk} of covariate j assigned to mixture component l.

5. For $j = 1, ..., J$; $l = 1, \dots L_j$ sample the mixture component means μ_{jl} from their normal posterior $\mathcal{N}(m_{jl}, M_{jl})$, where

$$M_{jl} = (N_{jl}/\psi_j + M_{0j}^{-1})^{-1},$$
$$m_{jl} = M_{jl}(N_{jl}\bar{\beta}_{jl}/\psi_j + M_{0j}^{-1} m_{0j})$$

and $\bar{\beta}_{jl}$ is the mean of all elements of $\boldsymbol{\beta}_j$ assigned to component l.

6. If a hyperprior is specified on the mixture component variances ψ_j, sample ψ_j for $j = 1, \dots, J$ from its inverse Gamma posterior $\mathcal{G}^{-1}(g_{jN}, G_{jN})$, where

$$g_{jN} = g_0 + c_j/2$$
$$G_{jN} = G_0 + \frac{1}{2} \sum_{k:S_{jk}=l} \sum_{l=0}^{L_j} (\beta_{jk} - \mu_{jl})^2.$$

7. Sample the vector of the latent allocation indicators $\boldsymbol{S}$ from the full conditional posterior

$$P(S_{jh} = l|\beta_{jh}, \boldsymbol{\mu}_j, \boldsymbol{\psi}_j) \propto \eta_{jl} f_{\mathcal{N}}(\beta_{jh}|\mu_{jl}, \psi_j) \qquad j = 1, \dots, J; h = 1, \dots, L_j$$

and update $\boldsymbol{b}_0(\boldsymbol{S}), \boldsymbol{B}_0(\boldsymbol{S}), N_{jl}$ and $\bar{\beta}_{jl}$ for $l = 1, \dots L_j$.

B Definitions

2.1 Silhouette coefficient

The silhouette coefficient in Rousseeuw (1987) is defined as follows. Let i be any object in the data set and A is the cluster to which it has been assigned. If cluster A

contains other objects apart from i, then $a(i)$ is the average dissimilarity of i to all other objects of A. $d(i, C)$ is the average dissimilarity of i to all objects in cluster C which represents any cluster different from A. Compute $d(i, C)$ for all clusters $C \neq A$ and denote by $b(i) = \min_{C \neq A} d(i, C)$. The silhouette coefficients is then computed as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)).}$$

2.2 Adjusted Rand index

The adjusted Rand index (Hubert and Arabie, 1985) is a form of the Rand index (Rand, 1971) which is adjusted for chance agreement. If n is the number of elements and $\mathbf{X} = \{X_1, X_2, ...X_r\}$ and $\boldsymbol{Y} = \{Y_1, Y_2, ..., Y_s\}$ are two clusterings of these elements, the adjusted Rand index is defined as

$$\mathrm{AR} = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

where a_i and b_j are the number of objects in X_i and Y_j, respectively and n_{ij} is the number of objects in $X_i \cap Y_j$.

C Further simulation results

3.1 Simulation results for covariates 1 to 3

We report simulation results for variables 1 to 3 of the simulation study in Tables 1 to 3.

	ν	freq	groups		AR		Error		FPR		FNR	
			most	pam	most	pam	most	pam	most	pam	most	pam
fixed	10	8324	2.8	3.0	0.89	0.99	0.07	0.00	0.00	0.01	0.07	0.00
	10^2	14466	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^3	14844	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^4	14847	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^5	14604	3.2	3.2	0.97	0.97	0.02	0.02	0.04	0.05	0.00	0.00
	10^6	13673	4.3	4.4	0.78	0.77	0.15	0.15	0.28	0.30	0.00	0.00
random	10	8054	2.8	3.0	0.90	0.99	0.07	0.00	0.00	0.00	0.06	0.00
	10^2	13940	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^3	14250	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^4	14343	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^5	14308	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^6	14308	3.0	3.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 1: Model selection results for variable 1, 10 categories, true number of groups is 3.

3.2 Parameter estimation accuracy and predictive performance

To evaluate the performance of the proposed approach with respect to estimation accuracy of the parameters we compute the mean squared error (MSE) of the coefficient estimates by averaging over all data set-specific mean squared errors

$$MSE^i = \frac{1}{C+1}((\beta_0^{true} - \hat{\beta}_0^i)^2 + \sum_{j=1}^{J}\sum_{k=1}^{c_j}(\beta_{jk}^{true} - \hat{\beta}_{jk}^i)'(\beta_{jk}^{true} - \hat{\beta}_{jk}^i)), \quad i = 1, \ldots, 100,$$

	ν	freq	groups		AR		Error		FPR		FNR	
			most	pam	most	pam	most	pam	most	pam	most	pam
fixed	10	14047	2.0	2.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^2	14601	2.0	2.0	1.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00
	10^3	14970	2.0	2.0	1.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00
	10^4	14929	2.0	2.0	0.99	0.99	0.00	0.00	0.00	0.01	0.00	0.00
	10^5	14263	2.4	2.5	0.75	0.70	0.10	0.13	0.17	0.21	0.00	0.00
	10^6	14095	3.4	3.5	0.30	0.28	0.33	0.35	0.52	0.54	0.00	0.00
random	10	13789	2.0	2.0	0.99	0.99	0.00	0.00	0.00	0.00	0.00	0.00
	10^2	13651	2.0	2.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^3	13711	2.0	2.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^4	13856	2.0	2.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^5	13915	2.0	2.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
	10^6	13678	2.0	2.0	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2: Model selection results for variable 2, 10 categories, true number of groups is 2.

where i is the number of the data set and $C = \sum_{j=1}^{J} c_j$ is the dimension of the vector of regression coefficients $\boldsymbol{\beta}$ in the full model.

In Figure 1 the MSE of the parameter estimates based on both model selection strategies as well as the MSE for the model averaged estimates ('av') are shown for different values of ψ_j, and fixed and random spike variances. For comparison, also the MSE of the penalized ML-estimates ('pen') and the estimates of the full model ('full') with a distinct effect for each level, and the true model ('true') with correctly fused levels,

	ν	freq	groups		AR		Error		FPR		FNR	
			most	pam	most	pam	most	pam	most	pam	most	pam
fixed	10	9154	1.1	2.0	0.86	0.00	0.01	0.10	0.03	0.20	-	-
	10^2	12591	1.1	2.1	0.90	0.00	0.01	0.14	0.02	0.26	-	-
	10^3	11897	1.8	2.0	0.26	0.00	0.23	0.28	0.33	0.41	-	-
	10^4	12003	3.1	3.3	0.00	0.00	0.47	0.50	0.63	0.67	-	-
	10^5	12402	4.9	4.6	0.00	0.00	0.63	0.64	0.82	0.81	-	-
	10^6	12709	7.1	5.5	0.00	0.00	0.74	0.68	0.92	0.85	-	-
random	10	9027	1.1	2.0	0.87	0.00	0.01	0.10	0.03	0.21	-	-
	10^2	10091	2.0	2.0	0.02	0.00	0.10	0.10	0.20	0.20	-	-
	10^3	10079	2.0	2.0	0.01	0.00	0.10	0.10	0.20	0.20	-	-
	10^4	10043	2.0	2.0	0.02	0.00	0.10	0.10	0.20	0.20	-	-
	10^5	10132	2.0	2.0	0.01	0.00	0.10	0.10	0.20	0.20	-	-
	10^6	10162	2.0	2.0	0.00	0.00	0.10	0.10	0.20	0.20	-	-

Table 3: Model selection results for variable 3, 10 categories, true number of groups is 1, i.e. all effects should be fused to the baseline.

both under a flat Normal prior, are shown.

For a fixed spike variance (plot on the left-hand side), the MSE of the selected models under both strategies is lower than for the full model and penalized regression. MSE is lowest for $\nu = 10^3$ and increases with larger ν, but never exceeds the MSE of the full model. Notable, the model averaged coefficient estimates ('av'), which do not rely on the selection of a specific model but rather average over all sampled models, outperform the full model estimates under a flat prior for each ν specification. Even

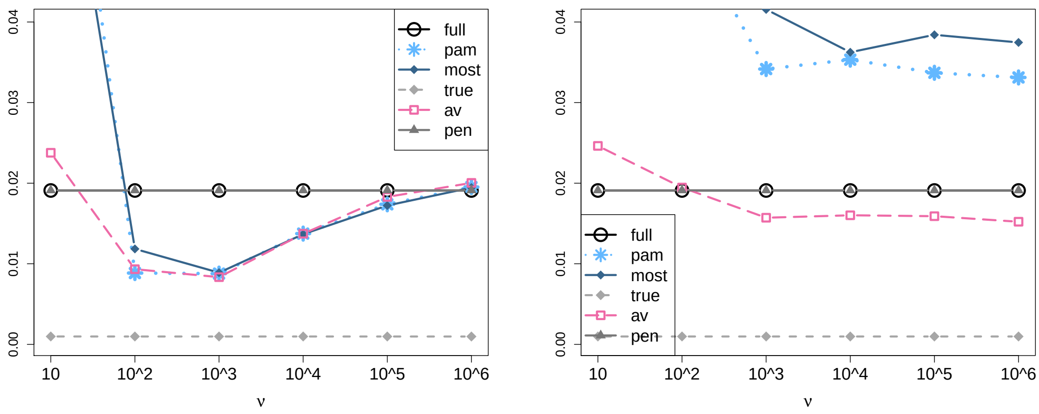

Figure 1: Simulation study: Mean squared error (MSE) of coefficient estimates for various values of ν with fixed component variance ψ_j (left) and hyperprior on ψ_j (right), averaged over 100 simulated data sets.

if the hyperprior on the the variance is specified (plot on the right-hand side) and the estimates of the selected models are worse than those of the full model (due to the sparse estimation of level groups in variable 4, see Table 3 in the main paper), the averaged estimates ('av') have smaller MSE than the full model. This indicates that the proposed mixture prior can also be used as an alternative to a non-informative prior in standard regression analysis, when just accurate parameter estimation and not model selection is the aim of the analysis, and more robust results in regard to prior specifications are desired.

Finally, to investigate the predictive performance of our approach, we generate 100 new data sets $(\boldsymbol{y}^{new}, \mathbf{X}^{new})$ with $N^{new} = 1,000$ observations and compute predictions of the response vector $\boldsymbol{y}^{new}$ based on $\mathbf{X}^{new}$ and the estimates of each of the 100 original data sets. The mean squared predictive error (MSPE) is computed as average of

$$MSPE^i = \frac{1}{N^{new}}(\boldsymbol{y}^{new} - \mathbf{X}^{new}\hat{\boldsymbol{\beta}}^i)'(\boldsymbol{y}^{new} - \mathbf{X}^{new}\hat{\boldsymbol{\beta}}^i)$$

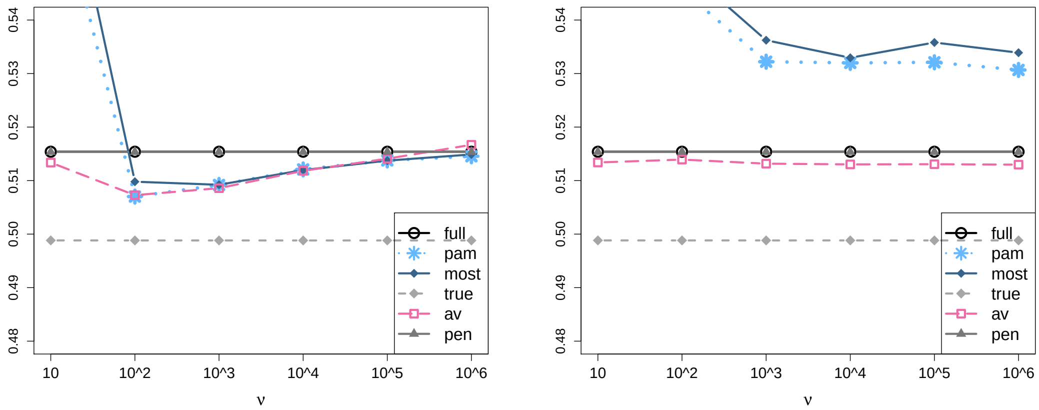

Figure 2: Simulation study: Mean squared prediction error (MSPE) of coefficient estimates for various values of ν with fixed component variance ψ_j (left) and hyperprior on ψ_j (right), averaged over 100 simulated data sets.

where $\hat{\boldsymbol{\beta}}^i$ is the estimate in data set i. The average MSPE is displayed in Figure 2. For fixed component variances, predictions from the selected models under the effect fusion prior ('most' and 'pam') and also using the model averaged estimates ('av') outperform those using the estimates from the full model and the regularised estimates ('pen'). However, if a hyperprior on the component variance is specified, the MSPE of the selected models is larger than the MSPE of the full model. Note that again model averaged estimates perform well yielding smaller prediction errors for values $\nu > 10$, thus outperforming estimates from the full model.

D SILC data

Table 4 describes the two-level classification scheme of the variable `job function` and the frequencies of the categories.

Level I (contract type)	Level II (skills)	Number of observations
apprentice	for white-collar worker	114
	for blue-collar worker	66
blue-collar worker	unskilled worker	143
	semi-skilled worker	413
	skilled worker	555
	foreman	83
white-collar worker	simple activities	85
	trained abilities/tasks	300
	medium abilities/tasks	543
	superior activities/tasks	388
	highly qualified activities	250
	leading activities	358
contract staff	simple activities	6
	craftsmanship activities	13
	auxiliary activities	8
	trained abilities/tasks	31
	medium abilities/tasks	57
	superior activities/tasks	61
	highly qualified or leading activities	21
officials	craftsmanship activities	10
	auxiliary activities	3
	trained abilities/tasks	27
	medium abilities/tasks	137
	superior activities/tasks	112
	highly qualified or leading activities	81
5	**25**	3865

Table 4: SILC data Austria 2010, variable `job function`: Five categories on the first level, 25 categories on the second level.

Table 5 describes the two-level classification scheme of the variable `economic sector` and the frequencies of the categories.

Level I	**Level II**	**Observations**
A Agriculture, forestry and fishing	A 01 Crop and animal production, hunting	20
	A 02 Forestry and logging	4
	A 03 Fishing and aquaculture	1
B Mining and quarrying	B 05 Mining of coal and lignite	-
	B 06 Extraction of crude petroleum, natural gas	2
	B 07 Mining of metal ores	1
	B 08 Other mining and quarrying	12
	B 09 Mining support service activities	-
C Manufacturing	C 10 Manufacture of food products	72
	C 11 Manufacture of beverages	11
	C 12 Manufacture of tobacco products	1
	C 13 Manufacture of textiles	12
	C 14 Manufacture of wearing apparel	10
	C 15 Manufacture of leather and related products	11
	C 16 Manufacture of wood; products of wood, cork	38
	C 17 Manufacture of paper and paper products	25
	C 18 Printing and reproduction of recorded media	24
	C 19 Manufacture of coke and refined petroleum products	4
	C 20 Manufacture of chemicals and chemical products	36
	C 21 Manufacture of basic pharmaceutical products	17
	C 22 Manufacture of rubber and plastic products	36
	C 23 Manufacture of other non-metallic mineral products	31
	C 24 Manufacture of basic metals	51
	C 25 Manufacture of fabricated metal products	102
	C 26 Manufacture of computer, electronic, optical products	34
	C 27 Manufacture of electrical equipment	46
	C 28 Manufacture of machinery and equipment	87

Level I	**Level II**	**Observations**
	C 29 Manufacture of motor vehicles	46
	C 30 Manufacture of other transport equipment	14
	C 31 Manufacture of furniture	29
	C 32 Other manufacturing	27
	C 33 Repair and installation of machinery	24
D Electricity, gas, steam supply	D 35 Electricity, gas, steam, air conditioning	31
E Water supply, waste management	E 36 Water collection, treatment and supply	4
	E 37 Sewerage	3
	E 38 Waste collection, materials recovery	12
	E 39 Remediation, other waste management services	-
F Construction	F 41 Construction of buildings	96
	F 42 Civil engineering	53
	F 43 Specialised construction activities	248
G Wholesale and retail trade	G 45 Wholesale, retail trade, repair of motor vehicles	87
	G 46 Wholesale trade	222
	G 47 Retail trade	253
H Transportation and storage	H 49 Land transport and transport via pipelines	99
	H 50 Water transport	1
	H 51 Air transport	5
	H 52 Warehousing and activities for transportation	81
	H 53 Postal and courier activities	37
I Accomodation and food service	I 55 Accommodation	74
	I 56 Food and beverage service activities	83
J Information and communication	J 58 Publishing activities	15
	J 59 Television , motion production, music recordings	6
	J 60 Programming and broadcasting activities	3
	J 61 Telecommunications	33
	J 62 Computer programming, consultancy	52

Level I	**Level II**	**Observations**
	J 63 Information service activities	7
K Finance and insurance	K 64 Financial service activities	119
	K 65 Insurance, reinsurance and pension funding	20
	K 66 Auxiliary to financial, insurance activities	27
L Real estate	L 68 Real estate activities	33
M Professional, scientific, technical act.	M 69 Legal and accounting activities	35
	M 70 Activities of head offices	11
	M 71 Architectural and engineering activitiess	62
	M 72 Scientific research and development	9
	M 73 Advertising and market research	17
	M 74 Other professional, scientific, technical activities	5
	M 75 Veterinary activities	1
N Administrative and support service	N 77 Rental and leasing activities	8
	N 78 Employment activities	29
	N 79 Travel agency, tour operator	17
	N 80 Security and investigation activities	7
	N 81 Services to buildings and landscape activities	31
	N 82 Office administrative, office support	13
O Public administration and defence	O 84 Public administration and defence	398
P Education	P 85 Education	289
Q Human health and social work	Q 86 Human health activities	189
	Q 87 Residential care activities	48
	Q 88 Social work activities without accommodation	39
R Arts, entertainment and recreation	R 90 Creative, arts and entertainment activities	10
	R 91 Libraries, archives, museums	7
	R 92 Gambling and betting activities	9
	R 93 Sports activities, amusement, recreation	15
S Other service activities	S 94 Activities of membership organisations	48
	S 95 Repair of computers, personal goods	2

Level I	**Level II**	**Observations**
	S 96 Other personal service activities	25
T Activities of household	T 97 Employers of domestic personnel	2
	T 98 Goods- and services-producing activities	-
U Activities of extraterritorial bodies	U 99 Activities of extraterritorial organisations	7
21	**84**	3865

Table 5: SILC data Austria 2010, variable `economic sector`: 21 categories on the first level, 84 categories on the second level.

References

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2** (1), 193–218.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.