The α–regression for compositional data: a unified framework for standard, spatially-lagged, spatial autoregressive and geographically-weighted regression models

Michail Tsagris[1] and Yannis Pantazis[2]

[1] Department of Economics, University of Crete, Greece
mtsagris@uoc.gr
[2] Institute of Applied & Computational Mathematics,
Foundation for Research & Technology - Hellas, Heraklion, Greece
pantazis@iacm.forth.gr

March 6, 2026

Abstract

Compositional data–vectors of non-negative components summing to unity–frequently arise in scientific applications where covariates influence the relative proportions of components, yet traditional regression approaches face challenges regarding the unit-sum constraint and zero values. This paper revisits the α–regression framework, which uses a flexible power transformation parameterized by α to interpolate between raw data analysis and log-ratio methods, naturally handling zeros without imputation while allowing data-driven transformation selection. We formulate α–regression as a non-linear least squares problem, study its asymptotic properties, provide efficient estimation via the Levenberg-Marquardt algorithm, and derive marginal effects for interpretation. The framework is extended to spatial settings through three models. The α–spatially lagged X regression model, which incorporates spatial spillover effects via spatially lagged covariates, with decomposition into direct and indirect effects. The α–spatially autoregressive regression model, that allows for spatial autocorrelation, and the geographically weighted α–regression, which allows coefficients to vary spatially for capturing local relationships. Applications to two real datasets illustrate the performance of the models and showcase that spatial extensions capture the spatial dependence and improve the predictive performance.

Keywords: compositional data, α–transformation, spatial regression

1 Introduction

Compositional data are characterized as vectors of non-negative components constrained to sum to a constant, conventionally normalized to unity. Compositional data structures arise across diverse scientific domains, as evidenced by the substantial body of methodological literature devoted to their rigorous statistical analysis[1]. The sample space of such data is defined by the standard simplex

$$\mathbb{S}^d = \left\{ (y_1, ..., y_D) \mid y_i \geq 0, \sum_{i=1}^{D} y_i = 1 \right\}, \tag{1}$$

where D denotes the dimensionality of the compositional vector, and $d = D - 1$ represents the degrees of freedom.

The methodological imperative to develop models specifically calibrated for compositional data has catalyzed considerable innovation, particularly in the contemporary statistical literature. The foundational framework was established by Aitchison (2003)–subsequently designated as Aitchison's model–predicated upon log-ratio transformations, thereby inaugurating the *log-ratio analysis* (LRA). This methodology was subsequently refined by Egozcue et al. (2003), who implemented an isometric log-ratio (ilr) transformation to preserve geometric properties. In contrast, the *stay-in-the-simplex approach* employs probability distributions and regression structures intrinsically defined on the simplex manifold. Notably, Dirichlet regression has been extensively utilized within compositional frameworks Gueorguieva et al. (2008), Hijazi and Jernigan (2009), Melo et al. (2009). Furthermore, Iyengar and Dey (2002) investigated the generalized Liouville distribution family, which accommodates negative or heterogeneous correlation structures, thereby extending beyond the restrictive positive correlation constraint of Dirichlet distributions. A less theoretically justified, yet occasionally employed strategy involves disregarding the unit-sum constraint and treating compositional data within a Euclidean framework—an approach designated as *raw data analysis* (RDA) (Baxter, 2001, Baxter et al., 2005). A fourth methodological paradigm employs the α–transformation family (Tsagris et al., 2011), which interpolates continuously between the RDA and LRA, thereby affording enhanced model flexibility while accommodating zero components naturally.

Given the ubiquity of compositional data across scientific disciplines, regression frameworks incorporating covariates have become methodologically essential. Applications encompass glacial sediment compositions, household expenditure allocations, geochemical soil analyses, morphometric measurements, electoral outcomes, pollution indices, and energy consumption patterns, among numerous other domains. The literature documents extensive applications of compositional regression methodology. For instance, Aitchison (2003) analyzed foraminiferal compositions across bathymetric gradients in oceanographic research. In hydrochemistry, Otero et al. (2005) employed regression techniques to discriminate anthropogenic from lithogenic sources of riverine contamination in Spain. Economic research, exemplified by Morais et al. (2018), has related market share dynamics to explanatory variables, while political scientists have modeled candidate vote proportions as functions of demographic and socioeconomic predictors (Katz and King, 1999). Compositional methodologies have similarly been deployed in bioinformatic analyses of microbiome community structures (Chen and Li, 2016, Shi et al., 2016, Xia et al., 2013).

A fundamental limitation of the aforementioned regression models concerns their inability to accommodate zero-valued components directly. Consequently, methodological developments

[1]For an extensive compilation of domain-specific applications involving compositional data, see (Tsagris and Stewart, 2020).

have addressed this constraint through various approaches. Scealy and Welsh (2011) proposed a transformation mapping compositional data to the unit hypersphere, introducing Kent regression which naturally accommodates structural zeros. From a Bayesian hierarchical perspective, Within econometric contexts, Mullahy (2015) formulated regression frameworks for fractional response data exhibiting zero inflation. Additional econometric strategies suitable for zero-augmented compositional data are systematically reviewed in Murteira and Ramalho (2016) and Alenazi (2022) who investigated the properties of ϕ-divergence regression models applicable to zero-inflated compositional data. Moreover, Tsagris (2015a) introduced a regression framework predicated upon minimization of the Jensen–Shannon divergence. Tsagris and Stewart (2018) extended the Dirichlet regression paradigm to accommodate zero components, yielding zero-adjusted Dirichlet regression.

Concerning spatial autocorrelation structures, the spatially lagged X (SLX) model represents a parsimonious specification incorporating spatial dependence exclusively through exogenous covariates, thereby excluding spatial lags of the dependent variable (Elhorst, 2014, LeSage and Pace, 2009). The spatial autoregressive (SAR) model (Kazar and Celik, 2012, Shi et al., 2025), analogous to temporal autoregressive processes, posits that observations are influenced by proximate spatial neighbors. Specifically, the SAR model expresses the dependent variable as a function of both explanatory covariates and a spatially weighted average of neighboring dependent variable realizations. Geographically weighted regression (GWR) constitutes a local regression methodology designed to capture spatially heterogeneous relationships (Brunsdon et al., 1996). In contrast to conventional regression, which assumes parameter stationarity, GWR permits spatial nonstationarity through location-specific coefficient estimation.

The integration of the spatial regression framework within the compositional data analysis represents a relatively narrow research area[2]. Leininger et al. (2013) synthesized hierarchical Bayesian models for zero-inflated compositional data, incorporating spatial random effects to accommodate local variation. Nguyen et al. (2021) and Yoshida et al. (2021) developed a SAR specification, and GWR, respectively, both employing the ilr transformation for compositional responses. Clarotto et al. (2022) introduced a novel power transformation, conceptually analogous to the α–transformation, specifically calibrated for geostatistical modeling of compositional data.

In this paper we adopt a pragmatic methodological stance, particularly tailored to regression with compositional data. The principal contribution of this paper is a unified framework for regression modelling of compositional data. We examine the α–regression (Tsagris, 2015b), that was proposed as a generalization of Aitchison's log-ratio regression (Aitchison, 2003), that naturally accommodates zero components, while offering flexibility via the α–transformation. First, we review the α–regression model, examine it as a non-linear least squares minimization problem and use a modified Levenberg-Marquardt algorithm, that is computationally efficient, to estimate the regression coefficients. We suggest two approaches to select the optimal value of α, and provide formulas for the marginal effects (MEs) of the covariates, including their asymptotic variance, We then establish the consistency and the asymptotic normality of the regression coefficients. Concluding the presentation α–regression, we discuss robust extensions

[2]The literature in spatial modelling contains more works, but most of them rely on log-ratio transformations prior to the application of standard spatial models.

and a simple method to incorporate compositional and Euclidean predictors.

The advantages of the α–regression are: a) ability to handle zeros naturally without imputation. b) Flexibility as α provides a continuum from power transforms to log-ratio methods. c) A predictive performance that is often higher compared to classical methods. d) A balance of the strengths of power transformations and log-ratio methods, providing a flexible and effective tool for predictive modeling on the simplex. A disadvantage though is the reduced interpretability of regression coefficients compared to log-ratio approaches.

We next extend the α–regression to accommodate spatial dependencies via three directions. The first extension is the α–SLX model, where we allow for spatial correlation in the predictors, that is we allow for spillover effect at the covariate level. The covariates affect directly the response, but also indirectly via the values of their neighbours. The second extension is the α–SAR model, where we place the correlation in the response side. The response is not only affected by the covariates, but also by the response values of the neighbours. Finally, we propose the GWαR model, where the regression coefficients are location specific. For all aforementioned spatial regression models, the selection of α and the free parameters is achieved via the spatial K-fold cross-validation (CV) protocol. Further, since the resulting regression coefficients are hard to interpret the effect of the covariates, we provide formulas to compute the MEs (and their asymptotic variance).

The next section discusses the α–regression, while section 3 generalizes this model to its spatial regression versions. Section 4 illustrates the performance of all regression models on two real datasets and Section 5 concludes the paper.

2 The α–regression

First the α–transformation, used for the α–regression, is defined, followed by the regression formulation.

2.1 The α–transformation

Tsagris et al. (2011) introduced the α–transformation, a power-based mapping designed for compositional data, $\mathbf{y} = (y_1, y_2, \ldots, y_D)$. For a given parameter $\alpha \in [-1, 1]$, the transformation is defined in two steps. Each component is raised to the power α and renormalized to remain in the simplex

$$\mathbf{u} = \left(\frac{y_i^\alpha}{\sum_{j=1}^D y_j^\alpha}, \ldots, \frac{y_D^\alpha}{\sum_{j=1}^D y_j^\alpha} \right). \tag{2}$$

This ensures $u = (u_1, \ldots, u_D)$ is itself a composition. To map compositions into Euclidean space for analysis, apply a linear transformation using the $D \times (D-1)$ Helmert sub-matrix $\mathbf{H}$:

$$\mathbf{y}_\alpha = \frac{1}{\alpha} \left(D\mathbf{u} - 1 \right) \mathbf{H}^\top, \tag{3}$$

where 1 denotes the D-dimensional vector of ones.

The α–transformation in (3) is a one-to-one transformation which maps data inside the simplex onto a subset of $\mathbb{R}^d$ and vice versa for $\alpha \neq 0$. Its corresponding sample space is

$$\mathbb{A}_\alpha^d = \left\{ \mathbf{H}\mathbf{w}_\alpha(\mathbf{y}) \mid -\frac{1}{\alpha} \leq w_{i,\alpha} \leq \frac{d}{\alpha}, \sum_{i=1}^d w_{i,\alpha} = 0 \right\}. \tag{4}$$

In effect, y_α resembles a Box–Cox style mapping, and the resulting y_α is an unconstrained vector in Euclidean space, suitable for standard multivariate statistical techniques. When $\alpha = 1$, the transformation corresponds (up to scaling) to raw data analysis (RDA). When $\alpha = -1$, the transformation is aligned with RDA as well, but using the inverse of the compositional data. At the limiting case, as $\alpha \to 0$, the transformation converges to the ilr transformation used in LRA.

$$\mathbf{y}_0 = \left(\log\left(\frac{y_1}{\prod_{j=1}^D x_j^{1/D}} \right), \ldots, \log\left(\frac{y_D}{\prod_{j=1}^D y_j^{1/D}} \right) \right) \mathbf{H}^\top. \tag{5}$$

Thus, the α–transformation provides a continuum between RDA and LRA, allowing analysts to choose the most appropriate representation of compositional data based on empirical performance or theoretical considerations.

2.2 The α–regression

The α–regression has the potential to improve the regression predictions with compositional data by adapting the α–transformation to the dataset's geometry. We assume that the conditional mean of the observed composition can be written as a non-linear function of some covariates

$$\mu_i = \begin{cases} \dfrac{1}{1 + \sum_{j=1}^D e^{\mathbf{x}^\top \beta_j}} & \text{for } i = 1 \\ \dfrac{e^{\mathbf{x}^\top \beta_i}}{1 + \sum_{j=1}^D e^{\mathbf{x}^\top \beta_j}} & \text{for } i = 2, \ldots, D \end{cases} \tag{6}$$

where

$$\boldsymbol{\beta}_i = \left(\beta_{0i}, \beta_{1i}, ..., \beta_{pi}\right)^\top, \ i = 1, ..., d \ \text{ and } p \text{ denotes the number of covariates.}$$

Tsagris (2015b) used the log-likelihood of the multivariate normal distribution, but in this paper the regression is formulated as a non-linear least squares problem, where the minimizing function is the sum of squares of the errors (SSE)

$$\text{SSE}\left(\mathbf{Y}, \mathbf{X}; \alpha, \mathbf{B}\right) = \sum_{i=1}^n \|\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\|_2^2 = \sum_{i=1}^n \left(\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\right)^\top \left(\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\right), \tag{7}$$

where $\mathbf{y}_{i,\alpha}$ and $\mathbf{m}_{i,\alpha}$ are the α–transformations applied to the i-th response and fitted compositional vectors, respectively and $\|.\|_2$ denotes the L_2 norm. Application of the stay-in-the-simplex power transformation (2) to the fitted vectors yields a simplified expression

$$\frac{\mu_i^\alpha}{\sum_{j=1}^D \mu_j^\alpha} = \frac{\left(\frac{e^{\mathbf{x}^\top \boldsymbol{\beta}_i}}{1+\sum_{j=1}^D e^{\mathbf{x}^\top \boldsymbol{\beta}_j}} \right)^\alpha}{\frac{1+\sum_{k=1}^D \left(e^{\mathbf{x}^\top \boldsymbol{\beta}_k}\right)^\alpha}{\left(1+\sum_{j=1}^D e^{\mathbf{x}^\top \boldsymbol{\beta}_j}\right)^\alpha}} = \frac{\left(e^{\mathbf{x}^\top \boldsymbol{\beta}_i}\right)^\alpha}{1 + \sum_{j=1}^D \left(e^{\mathbf{x}^\top \boldsymbol{\beta}_i}\right)^\alpha}.$$

For a given value of α, the matrix of the regression coefficients $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is estimated using a modification of the Levenberg-Marquardt algorithm[3]. The R package minpack.lm (Elzhov et al., 2023) is employed to this end[4].

2.2.1 Limiting case of $\alpha \to 0$

Tsagris et al. (2016) presented the proof that as $\alpha \to 0$, the α–transformation (3) converges to the ilr transformation (5). Following similar calculations one can show that

$$\lim_{\alpha \to 0} \frac{1}{\alpha} \left(D \frac{\mu_i^\alpha}{\sum_{j=1}^D \mu_j^\alpha} - 1 \right) \to \mathbf{x}\beta_i - \frac{\sum_{j=1}^D \mathbf{x}\beta_j}{D},$$

which corresponds to the regression after the centered log-ratio transformation [the ilr transformation (5) without the right multiplication by the Helmert matrix]. This implies that there are D vectors of β regression coefficients. But, since the set of regression coefficients sums to zero, if we subtract the first coefficient from the rest of the β_s we end up with the regression coefficients of the additive log-ratio (alr) regression

$$\log \left(\frac{y_i}{y_1} \right) = \mathbf{x}^\top \boldsymbol{\beta}_i, \quad i = 2, \ldots, D$$

2.2.2 Choosing α

In the regression setting the optimal value of α is data-driven, and there are two ways to estimate its value. The first is to minimize the Kullback-Leibler divergence (KLD) between the observed and fitted compositions $\text{KLD}\left(\mathbf{y}, \mu\right) = \sum_{i=1}^n \sum_{j=1}^D \mathbf{y}_{ij} \log y_{ij}/\mu_{ij}$. This results in a double minimization problem. For a given value of α one must minimize the SSE (7) in order to obtain the regression coefficients and then minimize the KLD with respect to α to obtain the optimal value of α. With the choice of the KLD, the value of α is independent of the SSE, since the SSE is not comparable across the different values of α. The second option is to examine α as a hyper-parameter whose value is chosen by minimizing the KLD via CV, e.g. 10-fold CV. (Tsagris, 2015b).

2.2.3 Marginal effects

To account for the difficult interpretation of the regression coefficients, the MEs, given below, may be used

$$\text{ME}_{ik} = \frac{\partial \mu_i}{\partial x_k} = \left\{ \begin{array}{ll} -\mu_1 \sum_{j=1}^d \beta_{jk}\mu_{j+1} & \text{for} \ \ i = 1 \\ \mu_i \left(\beta_{i-1,k} - \sum_{j=1}^d \beta_{jk}\mu_{j+1} \right) & \text{for} \ \ i = 2, \ldots, D \end{array} \right\}, \tag{8}$$

where $\sum_{i=1}^D \frac{\partial \mu_i}{\partial x_k} = 0$, because $\sum_{i=1}^D \mu_i = 1$. The sum of the MEs sums to zero, because if all components increase, one at least component must decrease by the same amount so that the unity sum constraint is preserved.

[3]This algorithm interpolates between the Gauss–Newton algorithm and the method of gradient descent.

[4]The relevant gradient vector, and the Hessian matrix are provided in the Appendix. The Newton-Raphson algorithm was implemented but exhibited slower convergence.

The average MEs (AME) across all observations are then computed as

$$\text{AME}_k = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial x_k}.$$

Standard errors can be computed via bootstrap or the delta method, accounting for estimation uncertainty in both $\widehat{\beta}$, $\widehat{\gamma}$, and $\widehat{\mu}$.

2.2.4 Standard error of the MEs

The covariance matrix of MEs is derived using the delta method. Let $\theta = \text{vec}(\mathbf{B})$ be the $dp \times 1$ vector of stacked regression coefficients, where $\mathbf{B} = (\beta_1, \ldots, \beta_d)$ is the $d \times p$ matrix of coefficients.

The general method of the Delta method reads that for a function $g(\theta)$ of the parameters:

$$\text{Var}(g(\theta)) \approx \mathbf{J}\text{Var}(\theta)\mathbf{J}^\top,$$

where $\mathbf{J} = \frac{\partial g}{\partial \theta^\top}$ is the Jacobian matrix.

For observation i, component ℓ, and covariate k, the ME is:

$$ME_{i\ell k} = m_{i\ell k} = \frac{\partial \mu_{i\ell}}{\partial x_{ik}}. \tag{9}$$

The Jacobian of the ME for the reference component ($\ell = 1$) is

$$\frac{\partial}{\partial \beta_{ms}}\left(\frac{\partial \mu_{i\ell}}{\partial x_{ik}}\right) = \begin{cases} -\delta_{sk}\mu_{i1}\mu_{im+1} - \frac{\partial \mu_{i1}}{\partial \beta_{ms}} \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1} - \mu_{i1}\sum_{j=1}^{d} \beta_{jk}\frac{\partial \mu_{ij+1}}{\partial \beta_{ms}} & \text{for } \ell = 1 \\ \delta_{sk}\delta_{m,\ell-1}\mu_{i\ell} - \delta_{sk}\mu_{i\ell}\mu_{im+1} + \frac{\partial \mu_{i\ell}}{\partial \beta_{ms}}\left[\beta_{\ell-1,k} - \sum_{j=1}^{d}\beta_{jk}\mu_{ij+1}\right] - & \\ \mu_{i\ell}\sum_{j=1}^{d}\beta_{jk}\frac{\partial \mu_{ij+1}}{\partial \beta_{ms}} & \text{for } \ell = 2, \ldots, D, \end{cases}$$

where δ_{sk} is the Kronecker delta ($\delta_{sk} = 1$ if $s = k$, 0 otherwise), and the derivatives of μ with respect to β are:

$$\frac{\partial \mu_{ij}}{\partial \beta_{rs}} = \begin{cases} -\mu_{i1}\mu_{ir+1}x_{is} & \text{if } j = 1 \\ \mu_{ij}(1-\mu_{ij})x_{is} & \text{if } j = r+1 \\ -\mu_{ij}\mu_{ir+1}x_{is} & \text{otherwise,} \end{cases}$$

where $r \in \{1, \ldots, d\}$ and $s \in \{1, \ldots, p\}$.

For observation i, component ℓ, and covariate k, the variance is

$$\text{Var}\left(\frac{\partial \mu_{i\ell}}{\partial x_{ik}}\right) = \mathbf{J}_{i\ell k}\,\text{Var}(\theta)\,\mathbf{J}_{i\ell k}^\top, \tag{10}$$

and the covariance matrix of the AMEs is

$$\text{Cov(AME)} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}_i\right)\text{Var}(\theta)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{J}_i\right)^\top. \tag{11}$$

2.3 Consistency and asymptotic normality

To establish the large-sample properties of the α-regression estimator $\hat{\theta}_n$, we impose the following assumptions or regularity conditions. The four structural assumptions delineated below are necessary to establish the consistency of the regression coefficients.

Assumption 2.1 (Data Generating Process)**.** The data $\{(y_i, x_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d.), and the conditional mean is correctly specified

$$\mathbb{E}[y_{i,\alpha} \mid x_i] = \mu_{i,\alpha}(\theta_0)\,.$$

Assumption 2.2 (Parameter Space)**.** The parameter space $\Theta \subset \mathbb{R}^{d(p+1)}$ is compact, and the true parameter $\theta_0 \in \mathrm{int}(\Theta)$.

Assumption 2.3 (Identification)**.** The population objective function

$$Q(\theta) = \mathbb{E}[\|y_{i,\alpha} - \mu_{i,\alpha}(\theta)\|^2]$$

has a unique minimum at θ_0. That is, for all $\theta \neq \theta_0$, $Q(\theta) > Q(\theta_0)$.

Assumption 2.4 (Continuity and Moment Conditions)**.** The function $\mu_{i,\alpha}(\theta)$ is continuous in θ for all $\theta \in \Theta$. Furthermore,

$$\mathbb{E}[\|y_{i,\alpha}\|^2] < \infty \quad \text{and} \quad \mathbb{E}[\sup_{\theta\in\Theta} \|\mu_{i,\alpha}(\theta)\|^2] < \infty\,.$$

Theorem 2.5 (Consistency)**.** *Under Assumptions 2.1–2.4,*

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \quad \textit{as} \quad n \to \infty\,.$$

Remark 2.6 (Necessity of Correct Specification in Consistency)*.* The assumption $\mathbb{E}[y_{i,\alpha} \mid x_i] = \mu_{i,\alpha}(\theta_0)$ (Assumption 2.1) ensures that the population objective function $Q(\theta) = \mathbb{E}[\|y_{i,\alpha} - \mu_{i,\alpha}(\theta)\|^2]$ is globally minimized at the true parameter θ_0. Under correct specification, the first-order condition at θ_0 becomes $\mathbb{E}[g_i(\theta_0)^\top(y_{i,\alpha} - \mu_{i,\alpha}(\theta_0))] = 0$. If the mean is misspecified, the estimator $\hat{\theta}_n$ will converge to a value θ^* that minimizes the approximation error, leading to asymptotic bias.

Remark 2.7 (Necessity of Bound on $y_{i,\alpha}$)*.* Although the raw composition y_i is bounded within the simplex $\mathcal{S}^{D-1}$, the transformed response $y_{i,\alpha}$ is not necessarily bounded. Specifically, for $\alpha < 0$, $y_{i,\alpha}$ involves terms of the form $y_{ij}^{-|\alpha|}$, which diverge as any component approaches zero. The assumption $\mathbb{E}[\|y_{i,\alpha}\|^2] < \infty$ is therefore essential to ensure that the objective function has a finite population expectation and that the variance of the residuals remains manageable.

Remark 2.8 (Dominance and Uniform Convergence)*.* The assumption $\mathbb{E}[\sup_{\theta\in\Theta} \|\mu_{i,\alpha}(\theta)\|^2] < \infty$ acts as a dominance condition. For the estimator $\hat{\theta}_n$ to be consistent, we require the sample objective $Q_n(\theta)$ to converge to the population objective $Q(\theta)$ uniformly over Θ. This moment condition ensures that the random objective function is bounded by an integrable function, satisfying the requirements for the USLLN.

Remark 2.9 (Identification and Compactness)*.* Consistency requires the identification of θ_0 as a well-separated minimum. By assuming Θ is compact and $\mu_{i,\alpha}(\theta)$ is continuous, the existence of a minimum is guaranteed. The moment conditions further ensure that the surface of the objective function is sufficiently smooth in expectation, preventing the estimator from converging to boundary points or spurious local minima as $n \to \infty$.

The next three assumptions are necessary to establish the asymptotic normality of the regression coefficients.

Assumption 2.10 (Smoothness)**.** The fitted values $\mu_{i,\alpha}(\theta)$ are twice continuously differentiable in θ on int(Θ). Let $g_i(\theta) = -\partial\mu_{i,\alpha}(\theta)/\partial\theta$ be the Jacobian.

Assumption 2.11 (Strengthened Moment Conditions)**.** The following hold: $\mathbb{E}[\sup_{\theta\in\Theta} \|g_i(\theta)\|^2] < \infty$ and
$\mathbb{E}[\sup_{\theta\in\Theta} \|\nabla^2 r_i(\theta)\|] < \infty$.

Assumption 2.12 (Non-singularity)**.** The Gram matrix $G(\theta_0) = \mathbb{E}[g_i(\theta_0)^\top g_i(\theta_0)]$ is positive definite.

Based on the regularity conditions established previously[cite: 11, 13, 17, 18], we now prove that the estimator $\hat{\theta}_n$ is asymptotically normally distributed.

Theorem 2.13 (Asymptotic Normality)**.** *Under Assumptions 2.1-2.12, as $n \to \infty$:*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, G(\theta_0)^{-1}\Omega(\theta_0)G(\theta_0)^{-1}) \tag{12}$$

where $G(\theta_0) = \mathbb{E}[g_i(\theta_0)^\top g_i(\theta_0)]$ and $\Omega(\theta_0) = \mathbb{E}[g_i(\theta_0)^\top r_i(\theta_0) r_i(\theta_0)^\top g_i(\theta_0)]$.

Remark 2.14 (Necessity of Correct Specification in Asymptotic Normality)*.* The assumption $\mathbb{E}[y_{i,\alpha} \mid x_i] = \mu_{i,\alpha}(\theta_0)$ (Assumption 2.1) is fundamental for the proof of Asymptotic Normality because the correct specification ensures that the second term of the Hessian, $\mathbb{E}[\nabla g_i(\theta_0)^\top r_i(\theta_0)]$, vanishes because $\mathbb{E}[r_i(\theta_0) \mid x_i] = 0$. Without this, the "bread" of the sandwich covariance matrix would require an additional complex term involving the second derivatives of the regression function, and the standard information matrix equality would fail.

2.4 Robust extensions of the α–regression

The α–regression is based upon minimization of the L_2 norm (7). In a similar fashion one may choose to minimize the sum of the absolute deviations, yielding the α–minimum absolute deviations (α–MAD) regression

$$\text{MAD}\left(\mathbf{Y}, \mathbf{X}; \alpha, \mathbf{B}\right) = \sum_{i=1}^{n}\sum_{j=1}^{d} |\mathbf{y}_{ij,\alpha} - \mu_{ij,\alpha}|.$$

Following Tsagris (2025b) the α–MAD regression was formulated as a univariate regression problem, by using the vectorization operation for the responses and by constructing the design matrix in a suitable manner. To make the estimation efficient the command `nlrq()` from the *R* package quantreg (Koenker et al., 2024) was utilized. This approach exhibits dependence on initialization and does not ensure convergence. Alternative models include alternative loss functions. Instead of the L_2 norm, one may use the L_1 norm $\sum_{i=1}^{n} \|\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\|_1$ leading to the α–spatial median regression[5]. Other options include Tukey's biweight loss function (Tukey, 1960), Hampel's loss function (Hampel, 1974) or Barron's general loss function (Barron, 2019), all available in the *R* package gslnls (Chau, 2025). A practical limitation of these approaches is the increased computational cost.

[5]One option for this regression is to use *R*'s built-in optimizers or iteratively reweighted least squares. The second option is computationally expensive as well, given the complexity of the derivatives involved.

2.5 The α–regression with continuous and compositional predictors

For convenience purposes, and without loss of generality, we will consider the case of a single composition, denoted by $\boldsymbol{Z}$. The composition $\boldsymbol{Z}$ is first transformed using the α–transformation (3), hence denoted by $\boldsymbol{Z}_\alpha$. Principal component analysis computes the eigenvectors $\boldsymbol{V}_\alpha$ of $\boldsymbol{Z}_\alpha$ and then the projections on to this orthonormal basis are computed, $\boldsymbol{S}_\alpha = \boldsymbol{Z}_\alpha \boldsymbol{V}_\alpha$. The extension of the α–regression to account for compositional predictors is straightforward.

The fitted values are given by

$$\mu_i^{\alpha'} = \begin{cases} \dfrac{1}{1+\sum_{j=1}^D e^{\boldsymbol{x}^\top \boldsymbol{\beta}_j + \boldsymbol{s}_{\alpha'}^\top \boldsymbol{\gamma}_j}} & \text{for } i = 1 \\ \dfrac{e^{\boldsymbol{x}^\top \boldsymbol{\beta}_i + \boldsymbol{s}_{\alpha'}^\top \boldsymbol{\gamma}_i}}{1+\sum_{j=1}^D e^{\boldsymbol{x}^\top \boldsymbol{\beta}_j + \boldsymbol{s}_{\alpha'}^\top \boldsymbol{\gamma}_j}} & \text{for } i = 2, \ldots, D. \end{cases} \tag{13}$$

The notation α' highlights that the value of α in the compositional predictors need not be the same as the one used when computing the SSE (7), and to clarify the difference, the new SSE may be written as

$$\text{SSE}\left(\boldsymbol{Y}, \boldsymbol{X}; \alpha, \alpha', \boldsymbol{B}\right) = \sum_{i=1}^n \|\boldsymbol{y}_{i,\alpha} - \boldsymbol{\mu}_{i,\alpha'}\|_2 = \sum_{i=1}^n \left(\boldsymbol{y}_{i,\alpha} - \boldsymbol{\mu}_{i,\alpha}^{\alpha'}\right)^\top \left(\boldsymbol{y}_{i,\alpha} - \boldsymbol{\mu}_{i,\alpha}^{\alpha'}\right). \tag{14}$$

Unlike the α–regression, the extension contains two independent α values, whose values can be chosen via the KLD as mentioned earlier.

3 The α–SLX, α–SAR and GWαR models

In the following sections we define three spatial extensions of the α–regression. The models presented cover the Euclidean predictors case, and the inclusion of compositional predictors is straightforward, and hence not covered.

3.1 The α–SLX model

The α–SLX model extends the standard α–regression by incorporating spatial spillover effects through the covariates. The fitted compositional values are given by:

$$\mu_i = \begin{cases} \dfrac{1}{1+\sum_{j=1}^D e^{\mathbf{x}^\top \beta_j + (\mathbf{Wx})^\top \gamma_j}} & \text{for } i = 1 \\ \dfrac{e^{\mathbf{x}^\top \beta_i + (\mathbf{Wx})^\top \gamma_i}}{1+\sum_{j=1}^D e^{\mathbf{x}^\top \beta_j + (\mathbf{Wx})^\top \gamma_j}} & \text{for } i = 2, \ldots, D. \end{cases} \tag{15}$$

The matrices of regression coefficients $\mathbf{B} = (\beta_1, \ldots, \beta_d)$ and $\boldsymbol{\Gamma} = (\gamma_1, \ldots, \gamma_d)$ are estimated in the same way as in the α–regression, and $\boldsymbol{W}$ is the contiguity matrix explained below.

3.1.1 The contiguity matrix

Some researchers tend to compute the Euclidean distance between two pairs of latitude and longitude, (ν_i, v_i) and (ν_j, v_j), $d_{ij} = \sqrt{(\nu_i - \nu_j)^2 + (v_i - v_j)^2}$. There is a fundamental flaw with this approach which is highlighted by Mardia and Jupp, 2000, pg. 13. Consider for instance the case of two coordinates whose latitude (or longitude) values are $359°$ and $1°$. Using the previous simplistic approach yields a distance between the two values $359° - 1° = 358°$, but the actual distance between them is only $2°$. To account for this, the pair of coordinates must first be transformed into their Euclidean coordinates, prior to the application of the Euclidean distance.

The locations (latitude and longitude) between a pair of observations, (ν_i, v_i) and (ν_j, v_j) are first mapped from their polar to their Cartesian coordinates (after transforming the degrees into radians)

$$\mathbf{c}_i = (\cos(\nu_i), sin(\nu_i)\cos(v_i), \sin(\nu_i)\sin(v_i)) \text{ and } \mathbf{c}_j = (\cos(\nu_j), sin(\nu_j)\cos(v_j), \sin(\nu_j)\sin(v_j)).$$

The Euclidean distance between $\mathbf{c}_i$ and $\mathbf{c}_j$ is

$$d(\mathbf{c}_i, \mathbf{c}_j) = d_{ij}^2 = \|\mathbf{c}_i - \mathbf{c}_j\|^2 = \|\mathbf{c}_i\|^2 + \|\mathbf{c}_j\|^2 - 2\mathbf{c}_i^\top \mathbf{c}_j = 2\left(1 - \mathbf{c}_i^\top \mathbf{c}_j\right).$$

For the i-th location, compute the region with the k nearest neighbors $\mathcal{C}_{ik}$ and zero the rest, that is

$$\tilde{w}_{ij} = \begin{cases} 1/d_{ij}^2 & \text{if } j \in \mathcal{C}_{ik} \\ \tilde{w}_{ij} = 0 & \text{else.} \end{cases} \tag{16}$$

The (i, j) elements of the contiguity matrix $\mathbf{W}$ are then defined as $w_{ij} = \tilde{w}_{ij} / \sum_{j=1}^n \tilde{w}_{ij}$.

3.1.2 Choosing α and k

The choice of the optimal values of α and of k is again data-driven and can be performed via CV, but this time the spatial 10-fold cross-validation CV protocol is employed, where the metric of performance is again the KLD.

The *R* package blockCV (Valavi et al., 2019) implements spatial cross-validation techniques designed to address the spatial autocorrelation inherent in geographical data. Unlike the traditional 10-fold CV, which can lead to overly optimistic model performance estimates when data points are spatially clustered (Roberts et al., 2017), the spatial version partitions data into spatially separated training and testing folds. This ensures that the testing data are spatially independent from the training data, providing more realistic assessments of model generalization to new geographic areas.

3.1.3 Spatial MEs

The spatial MEs (SMEs) consist of three components, the direct, the indirect and the total MEs. The following formulas are identical to the standard α–regression MEs (8), as they depend only on the β coefficients and do not involve spatial terms.

The direct SMEs measure the change in the covariate values

$$DSME_{ik} = \frac{\partial \mu_i}{\partial x_k} = \begin{cases} -\mu_1 \sum_{j=1}^d \beta_{jk}\mu_{j+1} & \text{for } i = 1 \\ \mu_i \left(\beta_{i-1,k} - \sum_{j=1}^d \beta_{jk}\mu_{j+1}\right) & \text{for } i = 2, \ldots, D. \end{cases}$$

The indirect (spillover) SMEs measure the impact of a change in the spatially lagged covariate $(\mathbf{Wx})_k$ (i.e., the weighted average of neighboring values) on the local composition component

μ_i. They have the same functional form as the direct effects, with γ replacing β. This structural symmetry reflects how spatial spillovers operate through the same multiplicative mechanism as direct effects.

$$IDSME_{ik} = \frac{\partial \mu_i}{\partial (\mathbf{W}\mathbf{x})_k} = \begin{cases} -\mu_1 \sum_{j=1}^{d} \gamma_{jk}\mu_{j+1} & \text{for } i = 1 \\ \mu_i \left(\gamma_{i-1,k} - \sum_{j=1}^{d} \gamma_{jk}\mu_{j+1}\right) & \text{for } i = 2, \ldots, D. \end{cases}$$

The total SMEs combine both direct and indirect SMEs representing the full impact of a simultaneous change in both local and neighboring covariate values.

$$TSME_{ik} = \frac{\partial \mu_i}{\partial x_k} + \frac{\partial \mu_1}{\partial (\mathbf{W}\mathbf{x})_k} = \begin{cases} -\mu_1 \sum_{j=1}^{d} (\beta_{jk} + \gamma_{jk})\mu_{j+1} & \text{for } i = 1 \\ \mu_i \left[(\beta_{i-1,k} + \gamma_{i-1,k}) - \sum_{j=1}^{d} (\beta_{jk} + \gamma_{jk})\mu_{j+1}\right] & \text{for } i = 2, \ldots, D. \end{cases}$$

3.1.4 Properties of the SMEs

Some properties regarding the SMEs are delineated below.

- The sum of the SMEs across all components equals zero:

$$\sum_{i=1}^{D} \frac{\partial \mu_i}{\partial x_k} = 0 \quad \text{and} \quad \sum_{i=1}^{D} \frac{\partial \mu_i}{\partial (\mathbf{W}\mathbf{x})_k} = 0$$

  This ensures that the composition remains on the simplex after perturbations.

- All SMEs depend on the current composition values μ, making them observation-specific and state-dependent.

- Direct and indirect effects share the same functional form, differing only in the coefficient vectors used (β vs γ).

- The contiguity matrix $\mathbf{W}$ determines which neighbors contribute to spillover effects. We remind that row-standardization is used such that $\sum_j w_{ij} = 1$.

- The standard error of the SMEs can be computed in a manner similar to the MEs of the α–regression and the formulas of their Jacobians can be found in Appendix F.

3.1.5 Prediction of new values

To predict the compositions for new observations $\mathbf{x}_{new}$, we must first construct the matrix $\mathbf{W}_{new}$ which contains the row normalized distances from the new locations to the existing ones, and then use the following formula

$$\widehat{\mu}_i = \begin{cases} \dfrac{1}{1 + \sum_{j=1}^{D} e^{\mathbf{x}_{new}^\top \beta_j + (\mathbf{W}_{new}\mathbf{x})^\top \gamma_j}} & \text{for } i = 1 \\ \dfrac{e^{\mathbf{x}_{new}^\top \beta_i + (\mathbf{W}_{new}\mathbf{x})^\top \gamma_i}}{1 + \sum_{j=1}^{D} e^{\mathbf{x}_{new}^\top \beta_j + (\mathbf{W}_{new}\mathbf{x})^\top \gamma_j}} & \text{for } i = 2, \ldots, D. \end{cases} \tag{17}$$

3.2 The α–SAR model

Inspired by the SAR for multinomial regression we define the following formulation

$$
\begin{aligned}
p_i &= \rho \mathbf{W} p_i + \mathbf{x}^\top \beta_i + \epsilon_i \\
\left(\mathbf{I}_n - \rho \mathbf{W}\right) p_i &= \mathbf{x}^\top \beta_i + \epsilon_i \\
p_i &= S\left(\rho\right)^{-1} \mathbf{x}^\top \beta_i + S\left(\rho\right)^{-1} \epsilon_i,
\end{aligned}
$$

for $i = 1, \ldots, D$, where $\rho \in (-1, 1)$ is the spatial autoregressive parameter measuring spillover strength, $S\left(\rho\right) = \mathbf{In} - \rho \mathbf{W}$ is the spatial multiplier matrix, and $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma_i^2 \mathbf{I}_n)$.

The fitted values are then defined in the same manner as in (6)

$$
\mu_i = \begin{cases} \dfrac{1}{1 + \sum_{j=1}^{D} e^{(\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{x}^\top \beta_j}} & \text{for } i = 1 \\ \dfrac{e^{(\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{x}^\top \beta_i}}{1 + \sum_{j=1}^{D} e^{(\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{x}^\top \beta_j}} & \text{for } i = 2, \ldots, D. \end{cases} \tag{18}
$$

Similarly to the α–regression, for a given value of α we minimize the SSE (7) in order to estimate the βs and the ρ parameter. The choice of α and k is again performed via the spatial 10-fold CV.

3.2.1 Computational challenges

The main obstacle faced during estimation of the α–SAR model is the inversion of the $n \times n$ matrix $S\left(\rho\right)$, a task that becomes computationally heavier as the sample size increases. Second, prior to performing the Levenberg-Marquardt algorithm we perform a grid search of ρ values, then estimate the parameters for a given value of ρ and choose the ρ that yields the minimum SSE. Each time, initial values for the βs are derived by the α–regression. Then, we use this ρ value and the resulting βs as starting values for the estimation of the model.

3.2.2 SMEs

The direct effects measure the impact of a change in location i's covariate on location i's own composition:

$$
DSME_{ilk} = \frac{\partial \mu_{i\ell}}{\partial x_{ik}} = \begin{cases} -\mu_{i1} \sum_{j=1}^{d} \beta_{jk} \mu_{ij+1} \cdot [S(\rho)^{-1}]_{ii} & \text{for } \ell = 1 \\ \mu_{i\ell} \left[\beta_{\ell-1,k} - \sum_{j=1}^{d} \beta_{jk} \mu_{ij+1}\right] \cdot [S(\rho)^{-1}]_{ii} & \text{for } \ell = 2, \ldots, D. \end{cases}
$$

The indirect effect at location i (summing spillovers from all neighbors) is

$$
IDSME_{iljk} = \frac{\partial \mu_{i\ell}}{\partial x_{jk}} = \begin{cases} -\mu_{i1} \sum_{j=1}^{d} \beta_{jk} \mu_{ij+1} \cdot \sum_{j \neq i} [S(\rho)^{-1}]_{ij} & \text{for } \ell = 1 \\ \mu_{i\ell} \left[\beta_{\ell-1,k} - \sum_{j=1}^{d} \beta_{jk} \mu_{ij+1}\right] \cdot \sum_{j \neq i} [S(\rho)^{-1}]_{ij} & \text{for } \ell = 2, \ldots, D \end{cases}
$$

The total SMEs are the sum of the direct and indirect effects.

3.2.3 Prediction of new values

Denote the new m covariate values by $\mathbf{X}^{new}$ located at new, unseen in the model, coordinates. We stack the new covariate values under the observed ones to create the augmented design matrix (Goulard et al., 2017)

$$\mathbf{X}^{aug} = \begin{pmatrix} \mathbf{X} & \mathbf{X}^{new} \end{pmatrix}.$$

Similarly define

$$\mathbf{W}^{aug} = \begin{pmatrix} \mathbf{W} & \mathbf{W}^{new} \\ \mathbf{W}^{new} & \mathbf{W} \end{pmatrix}$$

to be the augmented contiguity matrix, where $\mathbf{W}^{new}$ denotes the distances of the new locations from the observed ones. Note that $\mathbf{W}^{aug}$ is row standardised. $\mathbf{X}^{aug}$ contains $n+m$ rows, and $\mathbf{W}^{aug}$ is of dimensions $(n+m) \times (n+m)$, where n is the sample size of the observed sample, upon which the estimates are derived.

The predicted values are given by

$$\widehat{y}_{ij}^{aug} = \begin{cases} \dfrac{1}{1 + \sum_{\ell=1}^{D} e^{\left(\mathbf{I}_{n+m} - \rho \mathbf{W}^{aug}\right)^{-1} \left(\mathbf{x}_j^{aug}\right)^\top \beta_\ell}} & \text{for } i = 1 \\ \dfrac{e^{\left(\mathbf{I}_{n+m} - \rho \mathbf{W}^{aug}\right)^{-1} \left(\mathbf{x}_j^{aug}\right)^\top \beta_i}}{1 + \sum_{\ell=1}^{D} e^{\left(\mathbf{I}_{n+m} - \rho \mathbf{W}^{aug}\right)^{-1} \left(\mathbf{x}_j^{aug}\right)^\top \beta_\ell}} & \text{for } i = 2, \ldots, D. \end{cases}$$

Stacking the predicted values, in a matrix format, $\widehat{\mathbf{Y}}^{aug} = \begin{pmatrix} \widehat{\mathbf{Y}} \\ \widehat{\mathbf{Y}}^{new} \end{pmatrix}$, we observe that we the predictions for the new covariate values at the new locations are placed in the bottom m rows of $\widehat{\mathbf{Y}}^{aug}$.

3.3 The GWαR model

The GWαR model is a weighted α–regression scheme, but the difference is that the regression is performed n times, each time with different weights. The weighted SSE that must be minimized is

$$SSE\left(\mathbf{Y}, \mathbf{X}; \alpha, h, \mathbf{B}\right) = \sum_{i=1}^{n} \left(\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\right)^\top \mathbf{W}_i \left(\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\right), \tag{19}$$

where $\mathbf{W}_i = \text{diag}\left\{w_{i1}, \ldots, w_{in}\right\}$, is the weighting matrix corresponding to the weights allocated to each observation. A common weighting function is the Gaussian kernel

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2h^2}\right), \tag{20}$$

where d_{ij} is the distance between location i and j, and h is the bandwidth parameter controlling the degree of spatial smoothing.

As $\alpha \to 0$, the GWαR converges to the GWR after the alr transformation (Yoshida et al., 2021).

3.3.1 Choice of α and h

Choosing the optimal value of h in the classical GWR is typically achieved via the spatial 10-fold CV protocol, with the KLD acting as the metric of performance. The GWαR model entails an extra hyper-parameter, the α. This time the CV protocol is computationally more intensive. To alleviate the cost, the range of possible values α to be examined may be reduced, retaining only distinct values such as $\alpha = 0.1, 0.25, 0.5, 0.75, 1.0$. A heuristic approach to expedite the identification of the optimal α value involves performing the cross-validation protocol using the α–regression. However, empirical evidence suggests this strategy is inadvisable. Regarding the h hyper-parameter, following Gretton et al. (2012), Schrab et al. (2023) the median heuristic is employed as the starting point. This way, one knows a region to search for the optimal value of h.

3.3.2 Computational tricks to alleviate the computational burden

- Similarly to the α–regression, the stay-in-the-simplex power transformation (2) is written as

$$\frac{\mu_i^\alpha}{\sum_{j=1}^D \mu_j^\alpha} = \frac{\left(e^{\mathbf{x}^\top \boldsymbol{\beta}_i}\right)^\alpha}{1+\sum_{j=1}^D \left(e^{\mathbf{x}^\top \boldsymbol{\beta}_i}\right)^\alpha} = \frac{\left(e^{\alpha \mathbf{x}^\top \boldsymbol{\beta}_i}\right)}{1+\sum_{j=1}^D \left(e^{\alpha \mathbf{x}^\top \boldsymbol{\beta}_i}\right)}.$$

- The weighting function (20) becomes $w_{ij} = \exp\left(-\frac{d_{ij}^2}{2h^2}\right) = \exp\left(\frac{\mathbf{c}_i^\top \mathbf{c}_j - 1}{h^2}\right)$.

- The minimization of the SSE takes place for specific values of α and h. When passing the arguments of the SSE in the command `minpack.lm::nls.lm()`, the quantity $\alpha \mathbf{x}$ is pre-computed and passed as an argument.

- The function `minpack.lm::nls.lm()` requires a function that outputs the residuals. So, in order to perform weighted least squares we multiply the weights $\mathbf{w}_i$ by the residuals $\mathbf{r}_i$.

- For each observation i, we can compute the regression coefficients for different values of h. This is useful during the cross-validation protocol.

3.3.3 SMEs

The formula for the SMEs of the GWαR are nearly the same as those of the α–regression (8), but location specific

$$\frac{\partial \mu\left(\nu_i, v_i\right)}{\partial x_k} = \begin{cases} -\mu_1\left(\nu_i, v_i\right) \sum_{j=1}^d \beta_{jk}\left(\nu_i, v_i\right) \mu_{j+1}\left(\nu_i, v_i\right) & \text{for } i=1 \\ \mu_\ell\left(\nu_i, v_i\right) \left[\beta_{i-1,k}\left(\nu_i, v_i\right) - \sum_{j=1}^d \beta_{jk}\left(\nu_i, v_i\right) \mu_{j+1}\left(\nu_i, v_i\right)\right] & \text{for } \ell = 2, \ldots, D. \end{cases} \tag{21}$$

Just like in the α–regression, the $\sum_{\ell=1}^D \frac{\partial \mu_\ell(\nu_i, v_i)}{\partial x_k} = 0$, but this time, this is true for every location.

4 Application to real datasets

Real-data applications show that the α–regression can outperform the standard log-ratio-based regression, in terms of predictive performance, particularly when zeros are present, which can be further improved by taking into account the spatial dependencies.

The spatial 10-fold CV was employed to determine the values of the optimal hyper-parameters in each of the four regression models. To reduce the computational burden, 5 values for α were chosen, namely $\alpha = 0.1, 0.25, 0.5, 0.75, 1$. The values of k (for the α–SLX model) were set to $k = (2, \ldots, 15)$, and the bandwidth h of the GWαR was initially set equal to the median of the distances of the coordinates. A grid of 19 values spanning from $0.1h$ up to $10h$ were used for the GWαR model. The spatial 10-fold CV[6] was repeated 10 times[7] and the results were aggregated over these 10 times.

4.1 Agricultural economics dataset

Data regarding crop productivity in the Greek NUTS II region of Thessaly during the 2017-2018 cropping year were supplied by the Greek Ministry of Agriculture, also known as farm accountancy data network (FADN) data. The data refer to a sample of farms and initially they consisted of 20 crops, but after grouping and aggregation they were aggregated into five crop categories[8]. These crops are *Cereals*, *Cotton*, *Tree crops*, *Other annual crops and pasture* and *Grapes and wine*. For each of the 168 farms with unique coordinates, the cultivated area in each of these 5 grouped crops is known. Due to the presence of zero values the LRA approach, i.e. the family of the α–regression models presented earlier with $\alpha = 0$, is not applicable.

Figure 1(a) shows the location of Thessaly region in Greece, and Figure 1(b) shows the locations of the farms. The majority of the farms cultivate cereals and only a small proportion of farms cultivate grapes and wine. Specifically, 84.52% of the farms cultivate cereals, 50.00% cultivate Cotton, 40.48% maintain tree crops, 81.55% hold other annual crops and pasture, and finally only 16.67% of the farms own grapes and wine.

The goal is to examine the relationship between the composition of the cultivated area and the following covariates:

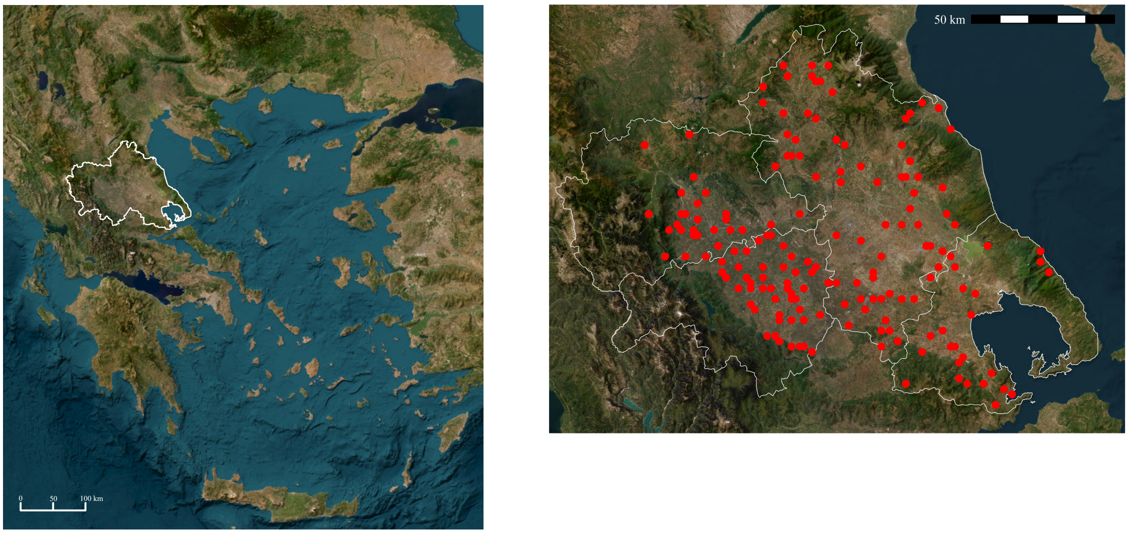
- Human Influence Index (HII, direct human influence on ecosystems). Zero value represents no human influence and 64 represents maximum human influence possible. The index uses all 8 measurements of human presence: Population Density/km^2, Score of Railroads, Score of Major Roads, Score of Navigable, Rivers, Score of Coastlines, Score of Nighttime Stable Lights Values, Urban Polygons, Land Cover Categories. The range of observed values is $16.08 - 46.69$, with an average of 29.021.
- The soil pH ($CaCl_2$). The range of values observed is between $0 - 6.99$ and the average is 6.33.
- Topsoil organic carbon content (SOC). The content (%) in the surface horizon of soils. The values ranged from 0.54 up to 10.07 with an average equal to 1.41.

Table 1 contains the results of the predictive performance estimation of the 4 models, aggregated over the 10 times repeated 10-fold spatial CV protocol. The α–SAR model exhibited

[6] The spatial CV was also applied to the α–regression to ensure a fair comparison.

[7] The time required to create the spatial folds was not accounted for.

[8] A larger version of this dataset was used in Mattas et al. (2025). Following the EU Regulation No1166/2008 that establishes a framework for European statistics at the level of agricultural holdings the aggregation took place across different output of crops.



(a) Region of Thessaly within Greece.

(b) The locations of the 168 farms.

Figure 1: The Thessaly region in Greece.

the optimal predictive performance, and, unexpectedly, the use of spatially lagged covariates deteriorated this performance. Two potential explanations that may account for this result are: a) the particular dataset may not exhibit strong spatial spillover effects and b) this suggests potential overfitting. In contrast, the GWαR model was the computationally most expensive among them.

Table 1: Agricultural economics dataset: average results regarding the optimal choice of α, k, h, KLD and running time (in seconds) for each of the four models.

Model	KLD	α	k	h	Running time
α–regression	0.810	0.775			3.308
α–SLX	1.603	0.550	6		172.860
α–SAR	0.608	1.000	5		1258.929
GWαR	0.869	0.675		3.369×10^{-3}	1233.419

We then fitted the regression models using the optimal parameters obtained based on the CV protocol and computed the correlations (component-wise) between the observed and fitted compositions. Table 2 contains these correlations. The α–SAR model has achieved the best fit. It is important to highlight that the spatial autoregressive parameter ρ of the α–SAR model was equal to -0.148 with a standard error equal to 0.0681. A t-test indicates that the coefficient is statistically significant.

4.2 Meuse river dataset

This dataset gives locations and topsoil heavy metal concentrations, along with a number of soil and landscape variables at the observation locations, collected in a flood plain of the river

Table 2: Agricultural economics dataset: Pearson correlations between each pair of the observed and fitted components for each of the four regression models.

	Cereals	Cotton	Tree crops	Other annual crops and pasture	Grapes and wine
α–regression	0.324	0.589	0.603	0.348	0.224
α–SLX	0.327	0.626	0.627	0.413	0.328
α–SAR	0.318	0.582	0.608	0.357	0.230
GWαR	0.497	0.741	0.760	0.485	0.372

Meuse, near the village of Stein (Netherlands). Heavy metal concentrations are from composite samples of a squared area of approximately 15m $\times$ 15m. There are measurements (all measured in mg kg^{-1} (ppm)): *topsoil cadmium concentration* (zero cadmium values in the original dataset have been shifted to 0.2 (half the lowest non-zero value)), *topsoil copper concentration*, *topsoil lead concentration* and *topsoil zinc concentration*. Figure 2 shows the map with locations of the sample. This dataset is characterized by the absence of zero values.

Figure 2: The flood plain of the river Meuse in Netherlands.

We have selected 3 covariates to associate the components with, namely the relative elevation above local river bed (in metres), the organic matter, kg (100 kg)$^{-1}$ soil (percent) and the distance to river Meuse (in metres), as obtained during the field survey.

Table 3 contains the average results regarding the optimal choice of α, k, h, KLD and running time (in seconds) for each of the four models. In this case the α–regression exhibits performance comparable to the α–SLX model, and the GWαR model exhibited the optimal performance, at the cost of duration. Moreover, the optimal α value did not remain consistent across models, and the value of zero, that corresponds to the ilr transformation (5), was never selected.

Table 3: Meuse river dataset: average results regarding the optimal choice of α, k, h, KLD and running time (in seconds) for each of the four models.

Model	KLD	α	k	h	Running time
α–regression	0.006	0.500			1.017
α–SLX	0.006	0.500	3		29.868
α–SAR	0.006	0.435	4		370.723
GWαR	0.036	0.250		590.523×10^{-6}	529.103

We then fitted the regression models using the optimal parameters obtained based on the CV protocol and computed the correlations (component-wise) between the observed and fitted compositions. Table 4 contains these correlations. The GWαR model has achieved the best fit. It is important to highlight that the spatial autoregressive parameter ρ of the α–SAR model was equal to -0.206 with a standard error equal to 0.164. Table 4 contains the correlations between each pair of the observed and fitted components for each of the four regression models. The α–SLX regression model has outperformed the rest.

Table 4: Meuse river dataset: Pearson correlations between each pair of the observed and fitted components for each of the four regression models.

	Cadmium	Copper	Lead	Zinc
α–regression	0.638	0.543	0.471	0.628
α–SLX	0.717	0.575	0.506	0.653
α–SAR	0.592	0.559	0.472	0.634
GWαR	0.648	0.558	0.482	0.638

5 Conclusions

The α–regression (Tsagris, 2015b) was revisited, and its theoretical properties were rigorously examined. A robustified version based on the quantile regression was proposed, although it is computationally intensive and may exhibit convergence difficulties. The inclusion of compositional predictors was facilitated via PCA, but this makes the regression computationally more challenging as it involves a second α parameter.

The α–regression was next expanded to account for spatial dependencies by introducing the α–SLX, the α–SAR and the GWαR models. For all models, formulas for the MEs were provided and their capabilities were tested on two real datasets.

The *R* package CompositionalSR (Tsagris, 2025a) was developed to perform all the α–regression modes, available to download from `CRAN`.

Future research could explore nonparametric spatially varying models for compositional data, as well as hybrid approaches that blend GWR with machine learning techniques for complex compositional systems. Further, a spatial non-linear autocorrelation test similar in the spirit of Moran's I test (Moran, 1950) would constitute a valuable direction for future research.

References

Aitchison, J. (2003). *The statistical analysis of compositional data.* New Jersey: Reprinted by The Blackburn Press.

Alenazi, A. (2022). f-divergence and parametric regression models for compositional data. *Pakistan Journal of Statistics and Operation Research*, 18(4):867–882.

Barron, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4331–4339.

Baxter, M. (2001). Statistical modelling of artefact compositional data. *Archaeometry*, 43(1):131–147.

Baxter, M., Beardah, C., Cool, H., and Jackson, C. (2005). Compositional data analysis of some alkaline glasses. *Mathematical Geology*, 37(2):183–196.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298.

Chau, J. (2025). *gslnls: GSL Multi-Start Nonlinear Least-Squares Fitting.* R package version 1.4.2.

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617.

Clarotto, L., Allard, D., and Menafoglio, A. (2022). A new class of α-transformations for the spatial analysis of compositional data. *Spatial Statistics*, 47:100570.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels.* SpringerBriefs in Regional Science. Springer, Heidelberg.

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., and Bolker, B. (2023). *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds.* R package version 1.2-4.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* John Wiley & Sons, Chichester, UK.

Goulard, M., Laurent, T., and Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, 12(2-3):304–325.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(1):723–773.

Gueorguieva, R., Rosenheck, R., and Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis*, 52(12):5344–5355.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hijazi, R. H. and Jernigan, R. W. (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability & Statistics*, 4(1):77–91.

Iyengar, M. and Dey, D. K. (2002). A semiparametric model for compositional data analysis in presence of covariates on the simplex. *Test*, 11(2):303–315.

Katz, J. and King, G. (1999). A statistical model for multiparty electoral data. *American Political Science Review*, pages 15–32.

Kazar, B. M. and Celik, M. (2012). *Spatial autoregression (SAR) model: Parameter estimation techniques*. Springer Science & Business Media.

Koenker, R. et al. (2024). *quantreg: Quantile Regression*. R package version 6.1.

Leininger, T. J., Gelfand, A. E., Allen, J. M., and Silander Jr, J. A. (2013). Spatial Regression Modeling for Compositional Data With Many Zeros. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):314–334.

LeSage, J. P. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC, Boca Raton, FL.

Mardia, K. and Jupp, P. (2000). *Directional Statistics*. London: John Wiley & Sons.

Mattas, K., Tsagris, M., and Tzouvelekas, V. (2025). Using synthetic farm data to estimate individual nitrate leaching levels. *American Journal of Agricultural Economics*, Forthcoming.

Melo, T. F., Vasconcellos, K. L., and Lemonte, A. J. (2009). Some restriction tests in a new class of regression models for proportions. *Computational Statistics & Data Analysis*, 53(12):3972–3979.

Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Using compositional and Dirichlet models for market share regression. *Journal of Applied Statistics*, 45(9):1670–1689.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Mullahy, J. (2015). Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods*, 4(1):71–100.

Murteira, J. M. R. and Ramalho, J. J. S. (2016). Regression analysis of multivariate fractional data. *Econometric Reviews*, 35(4):515–552.

Nguyen, T. H. A., Thomas-Agnan, C., Laurent, T., and Ruiz-Gazen, A. (2021). A simultaneous spatial autoregressive model for compositional data. *Spatial Economic Analysis*, 16(2):161–175.

Otero, N., Tolosana-Delgado, R., Soler, A., Pawlowsky-Glahn, V., and Canals, A. (2005). Relative vs. absolute statistical analysis of compositions: A comparative study of surface waters of a Mediterranean river. *Water Research*, 39(7):1404–1414.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Scealy, J. and Welsh, A. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society. Series B*, 73(3):351–375.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81.

Shi, J., Zhu, X., Zhou, J., Yu, B., and Wang, H. (2025). High-Dimensional Spatial Autoregression with Latent Factors by Diversified Projections. *Journal of the American Statistical Association*, pages 1–12.

Shi, P., Zhang, A., Li, H., et al. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040.

Tsagris, M. (2015a). A novel, divergence based, regression for compositional data. In *Proceedings of the 28th Panhellenic Statistics Conference, April, Athens, Greece.*

Tsagris, M. (2015b). Regression analysis with compositional data containing zero values. *Chilean Journal of Statistics*, 6(2):47–57.

Tsagris, M. (2025a). *CompositionalSR: Spatial Regression Models with Compositional Data.* R package version 1.0.

Tsagris, M. (2025b). Constrained least squares simplicial-simplicial regression. *Statistics and Computing*, 35(2):27.

Tsagris, M., Preston, S., and Wood, A. (2011). A data-based power transformation for compositional data. In *Proceedings of the 4rth Compositional Data Analysis Workshop, Girona, Spain.*

Tsagris, M., Preston, S., and Wood, A. T. (2016). Improved classification for compositional data using the α-transformation. *Journal of Classification*, 33(2):243–261.

Tsagris, M. and Stewart, C. (2018). A Dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39(3):398–412.

Tsagris, M. and Stewart, C. (2020). A folded model for compositional data analysis. *Australian & New Zealand Journal of Statistics*, 62(2):249–277.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., and Mann, H. B., editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 448–485. Stanford University Press, Stanford, CA.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232.

Xia, F., Chen, J., Fung, W. K., and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063.

Yoshida, T., Murakami, D., Seya, H., Tsutsumida, N., and Nakaya, T. (2021). Geographically weighted regression for compositional data: An application to the US household income compositions. In *Proceedings of the 11th International Conference on Geographic Information Science.*

Appendix

A Univariate spatial regression models

A.1 The SLX model

The SLX model provides a useful and interpretable framework for identifying spatial spillover effects through covariates alone. While it lacks the feedback mechanisms of models that include Wy (spatial autocorrelation of the dependent variable), it remains a robust and easily estimable tool for exploring spatial interactions. The structure of the SLX model allows researchers to capture how characteristics of neighboring spatial units affect local outcomes without introducing simultaneity. The general form of the SLX model is

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{k=1}^{p} \gamma_k \left(\sum_{j \neq i} w_{ij} x_{ik} \right) + \varepsilon_i, \tag{A.1}$$

where y denotes the dependent variable, x_k denotes the kt-h covariate, w_{ij} is the (i, j) element of the $n \times n$ spatial weights (contiguity) matrix $\mathbf{W}$ representing the spatial relationships between observations (e.g., contiguity or inverse distance), and $\sum_{j \neq i} w_{ij} x_k$ denotes the k-th spatially lagged covariate. The βs and γs are parameters corresponding to the direct (local) and indirect (spillover) effects, respectively, and ε is the classical error term.

A.2 The SAR model

The SAR model, which associates the response with its neighbours, may be written as

$$\begin{aligned} y_i &= \rho \mathbf{W} \mathbf{y} + \mathbf{X}_i \beta + \varepsilon_i \\ y_i &= S(\rho) \mathbf{X}_i \beta + S(\rho) \varepsilon_i, \end{aligned}$$

where $S(\rho) = (1 - \rho \mathbf{W})^{-1}$ and $\rho \in (0, 1)$ is the spatial autoregressive parameter and determines how much the dependent variable in one area is influenced by values in neighboring areas, according to the spatial structure defined by $\mathbf{W}$. In theory it can also take negative values, but positive values are more meaningful, in the sense that the neighbours of a farm that produces olive oil, for instance, affect the farm in a positive way.

The above model may also be written as $y_i = \mathbf{X}_i\gamma + u_i$ and be estimated using standard linear regression. The problem though is that we cannot disentangle the components of γ, we cannot separately identify the parameters $\boldsymbol{B}$ and ρ. Therefore, suitable estimation techniques must be employed identify them. Secondly, due to the endogeneity caused, the estimated regression coefficients are biased. Therefore suitable techniques must be employed, unless $\rho = 0$, i.e. no spatial dependence, in which case the standard linear regression suffices.

A.3 The GWR model

GWR has become a widely used technique in spatial statistics for modeling spatially varying relationships. Traditional regression assumes stationarity of relationships across space, but GWR relaxes this assumption by allowing coefficients to vary geographically (Brunsdon et al., 1996). Meanwhile, compositional data–datasets where variables represent proportions of a whole and are constrained to sum to unity–have gained attention in many disciplines, including environmental sciences, geology, and social sciences. When spatial heterogeneity and compositional constraints intersect, specialized methodological developments are required. The foundational work of Fotheringham et al. (2002) formalized GWR as a local regression technique that incorporates spatial weighting functions to account for the geographical location of observations.

The basic form of a standard multiple linear regression is:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i.$$

In GWR, the parameters are allowed to vary with location:

$$y_i = \beta_0(\nu_i, v_i) + \sum_{k=1}^{p} \beta_k(\nu_i, v_i)\, x_{ik} + \varepsilon_i,$$

where (ν_i, v_i) denotes the spatial coordinates of observation i (ν_i and v_i typically correspond to latitude and longitude, respectively), and $\beta_k(\nu_i, v_i)$ are the location-specific parameter estimates.

For each location (u_i, v_i), the parameter vector is estimated as:

$$\widehat{\beta}(\nu_i, v_i) = \left(X^\top W(\nu_i, v_i) X\right)^{-1} X^\top W(\nu_i, v_i)\mathbf{y},$$

where X is the design matrix and $W(\nu_i, v_i)$ is a spatial weighting matrix assigning higher weights to observations closer to (ν_i, v_i).

B Proof of Theorem 2.5

Proof. The proof proceeds in three steps.

Step 1: Uniform Convergence. Define the sample objective function

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \|y_{i,\alpha} - \mu_{i,\alpha}(\theta)\|^2 .$$

Using the inequality $|a-b|^2 \leq 2(|a|^2 + |b|^2)$ and Assumption 2.4,

$$\mathbb{E}[\|y_{i,\alpha} - \mu_{i,\alpha}(\theta)\|^2] \leq 2\Big(\mathbb{E}[\|y_{i,\alpha}\|^2] + \mathbb{E}\big[\sup_{\theta\in\Theta} \|\mu_{i,\alpha}(\theta)\|^2\big]\Big) < \infty .$$

Thus, $\|y_{i,\alpha} - \mu_{i,\alpha}(\theta)\|^2$ is integrable and by the Strong Law of Large Numbers (SLLN), for each fixed $\theta \in \Theta$, we have $Q_n(\theta) \xrightarrow{a.s.} Q(\theta)$ (i.e., point-wise). Since Θ is compact and $\|y_{i,\alpha} - \mu_{i,\alpha}(\theta)\|^2$ is continuous in θ (as a composition of continuous functions), and dominated by an integrable envelope as shown above, the Uniform Strong Law of Large Numbers (USLLN) implies

$$\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{a.s.} 0.$$

Step 2: Well-Separated Minimum. By Assumption 2.3 and the compactness of Θ, the minimum is well-separated. For any $\epsilon > 0$, there exists $\delta > 0$ such that:

$$\inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\epsilon} Q(\theta) \geq Q(\theta_0) + \delta. \tag{A.2}$$

Step 3: Consistency Argument. By definition, the estimator $\hat{\theta}_n = \arg\min_\theta Q_n(\theta)$ satisfies:

$$Q_n(\hat{\theta}_n) \leq Q_n(\theta_0). \tag{A.3}$$

For sufficiently large n and $\delta > 0$, Step 1 implies:

$$Q_n(\theta_0) \leq Q(\theta_0) + \frac{\delta}{2} \qquad \text{as well as} \qquad |Q_n(\hat{\theta}_n) - Q(\hat{\theta}_n)| < \frac{\delta}{2} . \tag{A.4}$$

Suppose, for contradiction, that $\left\|\hat{\theta}_n - \theta_0\right\| \geq \epsilon$ with positive probability. Then, for sufficiently large n:

$$\begin{aligned}
Q_n(\hat{\theta}_n) &\geq Q(\hat{\theta}_n) - |Q_n(\hat{\theta}_n) - Q(\hat{\theta}_n)| \\
&> Q(\hat{\theta}_n) - \frac{\delta}{2} && \text{(by (A.4))} \\
&\geq \inf_{\|\theta-\theta_0\|\geq\epsilon} Q(\theta) - \frac{\delta}{2} && \text{(by assumption)} \\
&\geq Q(\theta_0) + \delta - \frac{\delta}{2} && \text{(by (A.2))} \\
&\geq Q_n(\theta_0) && \text{(by (A.4))} .
\end{aligned}$$

which contradicts (A.3). Therefore,

$$\mathbb{P}(\|\hat{\theta}_n - \theta_0\| < \epsilon) \to 1 \quad \text{for all } \epsilon > 0,$$

which establishes $\hat{\theta}_n \xrightarrow{p} \theta_0$. □

C Proof of Theorem 2.13

Based on the regularity conditions established previously[cite: 11, 13, 17, 18], we now prove that the estimator $\hat{\theta}_n$ is asymptotically normally distributed.

Theorem C.1 (Asymptotic Normality)**.** *Under Assumptions 1–7[cite: 11, 12, 13, 17, 18, 19], as $n \to \infty$:*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, G(\theta_0)^{-1}\Omega(\theta_0)G(\theta_0)^{-1}) \tag{A.5}$$

where $G(\theta_0) = \mathbb{E}[g_i(\theta_0)^\top g_i(\theta_0)]$ [cite: 18] and $\Omega(\theta_0) = \mathbb{E}[g_i(\theta_0)^\top r_i(\theta_0) r_i(\theta_0)^\top g_i(\theta_0)]$.

Proof. The proof relies on a Taylor expansion of the first-order condition (FOC) of the objective function $Q_n(\theta)$ [cite: 9, 10].

Step 1: First-Order Condition. By Theorem 1, $\hat{\theta}_n \xrightarrow{p} \theta_0$[cite: 21]. Since $\theta_0 \in \text{int}(\Theta)$[cite: 11], for sufficiently large n, the estimator $\hat{\theta}_n$ must satisfy the FOC:

$$\nabla Q_n(\hat{\theta}_n) = \frac{2}{n}\sum_{i=1}^{n} g_i(\hat{\theta}_n)^\top r_i(\hat{\theta}_n) = 0 \tag{A.6}$$

where $r_i(\theta) = y_{i,\alpha} - \mu_{i,\alpha}(\theta)$ [cite: 9] and $g_i(\theta) = -\frac{\partial \mu_{i,\alpha}(\theta)}{\partial \theta}$[cite: 15].

Step 2: Taylor Expansion. Applying the Mean Value Theorem to $\nabla Q_n(\hat{\theta}_n)$ around the true parameter θ_0:

$$0 = \nabla Q_n(\theta_0) + \nabla^2 Q_n(\bar{\theta})(\hat{\theta}_n - \theta_0) \tag{A.7}$$

where $\bar{\theta}$ lies on the line segment connecting $\hat{\theta}_n$ and θ_0. Rearranging to solve for the scaled difference:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -[\nabla^2 Q_n(\bar{\theta})]^{-1}\sqrt{n}\nabla Q_n(\theta_0) \tag{A.8}$$

Step 3: Convergence of the Hessian (The Bread). The Hessian of the objective function is given by:

$$\nabla^2 Q_n(\theta) = \frac{2}{n}\sum_{i=1}^{n}\Big(g_i(\theta)^\top g_i(\theta) + \nabla g_i(\theta)^\top r_i(\theta)\Big) \tag{A.9}$$

As $n \to \infty$, $\bar{\theta} \xrightarrow{p} \theta_0$. By the Uniform Law of Large Numbers and the fact that $\mathbb{E}[r_i(\theta_0) \mid x_i] = 0$[cite: 19, 20]:

$$\nabla^2 Q_n(\bar{\theta}) \xrightarrow{p} 2\mathbb{E}[g_i(\theta_0)^\top g_i(\theta_0)] = 2G(\theta_0) \tag{A.10}$$

where $G(\theta_0)$ is positive definite[cite: 18].

Step 4: Convergence of the Gradient (The Meat). The term $\sqrt{n}\nabla Q_n(\theta_0)$ is a sum of i.i.d. random variables[cite: 19]:

$$\sqrt{n}\nabla Q_n(\theta_0) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n} g_i(\theta_0)^\top r_i(\theta_0) \tag{A.11}$$

By the Central Limit Theorem, since $\mathbb{E}[g_i(\theta_0)^\top r_i(\theta_0)] = 0$[cite: 20]:

$$\sqrt{n}\nabla Q_n(\theta_0) \xrightarrow{d} N(0, 4\Omega(\theta_0)) \tag{A.12}$$

where $\Omega(\theta_0) = \mathbb{E}[g_i(\theta_0)^\top r_i(\theta_0) r_i(\theta_0)^\top g_i(\theta_0)]$.

Step 5: Final Result via Slutsky's Theorem. Combining the results using Slutsky's Theorem:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &\xrightarrow{d} -[2G(\theta_0)]^{-1}N(0, 4\Omega(\theta_0)) \\ &= N(0, [2G(\theta_0)]^{-1}(4\Omega(\theta_0))[2G(\theta_0)]^{-1}) \\ &= N(0, G(\theta_0)^{-1}\Omega(\theta_0)G(\theta_0)^{-1})\end{aligned}$$

This completes the proof of asymptotic normality with the sandwich covariance matrix. □

D Gradient vector and Hessian matrix for the α–regression

The least squares objective function is

$$\mathrm{SSE}(\mathbf{Y}, \mathbf{X}; \alpha, \mathbf{B}) = -\frac{1}{2}\mathrm{tr}[(\mathbf{y}_\alpha - \mu_\alpha)^\top(\mathbf{y}_\alpha - \mu_\alpha)],$$

where $\mathbf{y}_\alpha$ is the α–transformed observed compositional data ($n \times d$ matrix), μ_α is the α–transformed fitted compositional values ($n \times d$ matrix), n is the number of observations, and $d = D - 1$ where D is the number of components in the composition.

The fitted compositional values come from the inverse alr transformation:

$$\mu_1 = \frac{1}{1 + \sum_{j=1}^{d} e^{x^\top \beta_j}}, \quad \mu_i = \frac{e^{x^\top \beta_{i-1}}}{1 + \sum_{j=1}^{d} e^{x^\top \beta_j}}, \quad i = 2, \ldots, D.$$

D.1 The α–transformation

The α–transformation consists of two steps:

Step 1: Power transformation

$$u_i = \frac{\mu_i^\alpha}{\sum_{j=1}^{D} \mu_j^\alpha}, \quad i = 1, \ldots, D.$$

Step 2: Helmert transformation

$$z = \frac{1}{\alpha} H(Du - j_D),$$

where H is the $d \times D$ Helmert sub-matrix and j_D is a D-dimensional vector of ones.

D.2 First Derivatives (Gradient)

D.2.1 Main Gradient Formula

$$\frac{\partial l(\alpha)}{\partial \beta_k} = \mathrm{tr}\left[(\mathbf{y}_\alpha - \mu_\alpha)^\top \frac{\partial \mu_\alpha}{\partial \beta_k}\right].$$

D.2.2 Expanded Gradient Formula

$$\frac{\partial l(\alpha)}{\partial \beta_k} = \sum_{i=1}^{n}\sum_{m=1}^{d}\sum_{\ell=1}^{D}\sum_{p=1}^{D} r_{\alpha,im} \cdot \frac{D}{\alpha} H_{m\ell} \cdot \frac{\partial u_{i\ell}}{\partial \mu_{ip}} \cdot \frac{\partial \mu_{ip}}{\partial \beta_k} \cdot x_i,$$

where $r_{\alpha,im} = y_{\alpha,im} - m_{\alpha,im}$ are the residuals in α–transformed space, $H_{m\ell}$ is the (m, ℓ) element of the Helmert sub-matrix, and x_i is the covariate vector for observation i.

D.2.3 Jacobian of Power Transformation

$$\frac{\partial u_{i\ell}}{\partial \mu_{ip}} = \begin{cases} \dfrac{\alpha \mu_{i\ell}^{\alpha-1}}{\sum_{j=1}^{D} \mu_{ij}^{\alpha}} \left(1 - \dfrac{\mu_{i\ell}^{\alpha}}{\sum_{j=1}^{D} \mu_{ij}^{\alpha}}\right) & \text{if } \ell = p \\ [3ex] - \dfrac{\alpha \mu_{i\ell}^{\alpha} \mu_{ip}^{\alpha-1}}{(\sum_{j=1}^{D} \mu_{ij}^{\alpha})^2} & \text{if } \ell \neq p \end{cases}.$$

Let $T_i = \sum_{j=1}^{D} \mu_{ij}^{\alpha}$. In compact form:

$$\frac{\partial u_{i\ell}}{\partial \mu_{ip}} = \frac{\alpha \mu_{ip}^{\alpha-1}}{T_i} \left(\delta_{\ell p} - \frac{\mu_{i\ell}^{\alpha}}{T_i}\right),$$

where $\delta_{\ell p}$ is the Kronecker delta.

D.2.4 Jacobian of Multinomial Logit

Let $S_i = 1 + \sum_{j=1}^{d} e^{x_i^\top \beta_j}$.

$$\frac{\partial \mu_{ip}}{\partial \beta_k} = \begin{cases} -\mu_{i1}\mu_{ik} x_i & \text{if } p = 1 \\ \mu_{ik}(1 - \mu_{ik}) x_i & \text{if } p = k+1 \\ -\mu_{ip}\mu_{ik} x_i & \text{if } p \neq 1, p \neq k+1 \end{cases},$$

where $\mu_{ik} = \mu_{i,k+1}$ (the $(k+1)$-th component of the composition).

D.2.5 Vectorized Gradient Formula

$$\frac{\partial l(\alpha)}{\partial \beta_k} = X^\top w_k,$$

where the weight vector $w_k \in \mathbb{R}^n$ has elements:

$$w_{k,i} = \left\{ r_{\alpha,i}^\top \cdot \frac{D}{\alpha} H \cdot J_u(i) \cdot J_\mu(i,k) \right\}.$$

Diagonal Contribution

$$w_{k,i}^{\text{diag}} = \sum_{\ell=1}^{D} r_{\alpha,i\ell} H_\ell J_{u,\text{diag}}(i,\ell) J_\mu(i,\ell,k)$$

where $J_{u,\text{diag}}(i,\ell) = \frac{\alpha \mu_{i\ell}^{\alpha-1}}{T_i} \left(1 - \frac{\mu_{i\ell}^{\alpha}}{T_i}\right)$.

Off-Diagonal Contribution

$$w_{k,i}^{\text{off-diag}} = -\frac{\alpha}{T_i^2} \left[\left(\sum_{\ell=1}^{D} r_{\alpha,i\ell} H_\ell \mu_{i\ell}^{\alpha} \right) \left(\sum_{p=1}^{D} \mu_{ip}^{\alpha-1} J_\mu(i,p,k) \right) - \sum_{\ell=1}^{D} r_{\alpha,i\ell} H_\ell \mu_{i\ell}^{\alpha} \mu_{i\ell}^{\alpha-1} J_\mu(i,\ell,k) \right].$$

Total Weight:

$$w_{k,i} = w_{k,i}^{\text{diag}} + w_{k,i}^{\text{off-diag}}.$$

D.3 Hessian matrix for the α–regression

The sum of squares of the errors is:

$$l(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{\alpha,ij} - m_{\alpha,ij})^2.$$

We will compute the Hessian matrix including all second-order terms. The gradient is

$$\frac{\partial l(\alpha)}{\partial \beta_k} = \sum_{i=1}^{n}\sum_{j=1}^{d} r_{\alpha,ij}\frac{\partial m_{\alpha,ij}}{\partial \beta_k},$$

where $r_{\alpha,ij} = y_{\alpha,ij} - m_{\alpha,ij}$. The structure of the Hessian matrix is:

$$H_{\text{exact}} = \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,d} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ H_{d,1} & H_{d,2} & \cdots & H_{d,d} \end{bmatrix}.$$

Each block $H_{k,k'} \in \mathbb{R}^{p\times p}$ includes both first and second-order terms.

The derivative with respect to $\beta_{k'}$ is:

$$\frac{\partial^2 l(\alpha)}{\partial \beta_k \partial \beta_{k'}} = \underbrace{-\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\partial m_{\alpha,ij}}{\partial \beta_k}\frac{\partial m_{\alpha,ij}}{\partial \beta_{k'}}}_{\text{First-order term (GN)}} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{d} r_{\alpha,ij}\frac{\partial^2 m_{\alpha,ij}}{\partial \beta_k \partial \beta_{k'}}}_{\text{Second-order term}}.$$

D.3.1 First-Order Term (Gauss-Newton Part)

This is identical to the Gauss-Newton approximation:

$$H^{(1)}_{k,k'} = -\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\partial m_{\alpha,ij}}{\partial \beta_k}\frac{\partial m_{\alpha,ij}}{\partial \beta_{k'}} = -X^\top \text{diag}(W_{k,k'})X.$$

where

$$W_{k,k'}(i,i) = \sum_{j=1}^{d}\frac{\partial m_{\alpha,ij}}{\partial \beta_k}\cdot\frac{\partial m_{\alpha,ij}}{\partial \beta_{k'}}.$$

D.3.2 Second-Order Term (Exact Correction)

Computation of $\frac{\partial^2 m_{\alpha,ij}}{\partial \beta_k \partial \beta_{k'}}$.

D.3.3 Chain Rule for Second Derivative

The chain rule for first derivative is

$$\frac{\partial m_{\alpha,i}}{\partial \beta_k} = \frac{D}{\alpha}H \cdot J_u(i)\cdot J_\mu(i,k).$$

Taking the derivative with respect to $\beta_{k'}$:

$$\frac{\partial^2 m_{\alpha,i}}{\partial \beta_k \partial \beta_{k'}} = \frac{D}{\alpha}H\cdot\left[\frac{\partial J_u(i)}{\partial \beta_{k'}}\cdot J_\mu(i,k) + J_u(i)\cdot\frac{\partial J_\mu(i,k)}{\partial \beta_{k'}}\right].$$

D.3.4 Second Derivative of Power Transformation

We need $\frac{\partial J_u(i)}{\partial \beta_{k'}}$, which involves $\frac{\partial^2 u_\ell}{\partial \mu_p \partial \mu_q}$. Let $T_i = \sum_{j=1}^{D} \mu_{ij}^\alpha$.

Diagonal-Diagonal: $\ell = p = q$

$$\frac{\partial^2 u_\ell}{\partial \mu_\ell^2} = \frac{\alpha(\alpha-1)\mu_\ell^{\alpha-2}}{T}\left(1 - \frac{\mu_\ell^\alpha}{T}\right) - \frac{2\alpha^2 \mu_\ell^{2\alpha-2}}{T^2} + \frac{2\alpha^2 \mu_\ell^{3\alpha-2}}{T^3}.$$

Diagonal-Off-diagonal: $\ell = p \neq q$

$$\frac{\partial^2 u_\ell}{\partial \mu_\ell \partial \mu_q} = -\frac{\alpha(\alpha-1)\mu_\ell^{\alpha-1}\mu_q^{\alpha-1}}{T^2}\left(1 - \frac{\mu_\ell^\alpha}{T}\right) - \frac{\alpha^2 \mu_\ell^{\alpha}\mu_q^{\alpha-1}}{T^2} + \frac{2\alpha^2 \mu_\ell^{2\alpha-1}\mu_q^{\alpha-1}}{T^3}.$$

Off-diagonal-Off-diagonal: $\ell \neq p, \ell = q$

$$\frac{\partial^2 u_\ell}{\partial \mu_\ell \partial \mu_p} = -\frac{\alpha(\alpha-1)\mu_\ell^{\alpha-1}\mu_p^{\alpha-1}}{T^2}\left(1 - \frac{\mu_\ell^\alpha}{T}\right) - \frac{\alpha^2 \mu_\ell^{2\alpha-1}\mu_p^{\alpha-1}}{T^2} + \frac{2\alpha^2 \mu_\ell^{3\alpha-1}\mu_p^{\alpha-1}}{T^3}.$$

Fully Off-diagonal: $\ell \neq p, \ell \neq q, p \neq q$

$$\frac{\partial^2 u_\ell}{\partial \mu_p \partial \mu_q} = -\frac{\alpha(\alpha-1)\mu_\ell^{\alpha}\mu_p^{\alpha-2}\delta_{pq}}{T^2} - \frac{\alpha^2 \mu_\ell^{\alpha}\mu_p^{\alpha-1}\mu_q^{\alpha-1}}{T^2} + \frac{2\alpha^2 \mu_\ell^{\alpha}\mu_p^{\alpha-1}\mu_q^{\alpha-1}}{T^3},$$

where δ_{pq} is the Kronecker delta.

D.3.5 General Formula for Hessian of Power Transformation

Let $H_u(i, \ell)$ denote the $D \times D$ Hessian matrix for component ℓ of u_i:

$$[H_u(i,\ell)]_{pq} = \frac{\partial^2 u_{i\ell}}{\partial \mu_{ip} \partial \mu_{iq}}.$$

Then:

$$\frac{\partial J_u(i)}{\partial \beta_{k'}} = \sum_{\ell=1}^{D}\sum_{p=1}^{D} H_u(i,\ell)_{pq} \cdot \frac{\partial \mu_{ip}}{\partial \beta_{k'}} \cdot e_\ell e_q^\top$$

where e_ℓ is the ℓ-th standard basis vector. This becomes a $D \times D$ matrix where each element is:

$$\left[\frac{\partial J_u(i)}{\partial \beta_{k'}}\right]_{\ell q} = \sum_{p=1}^{D} \frac{\partial^2 u_{i\ell}}{\partial \mu_{ip} \partial \mu_{iq}} \cdot \frac{\partial \mu_{ip}}{\partial \beta_{k'}}.$$

D.3.6 Second Derivative of Multinomial Logit

We need $\frac{\partial J_\mu(i,k)}{\partial \beta_{k'}}$, which involves $\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \beta_{k'}}$. Let $\mu_{ik} = \mu_{i,k+1}$ and $\mu_{ik'} = \mu_{i,k'+1}$.

For component $p = 1$ (reference):

$$\frac{\partial^2 \mu_{i1}}{\partial \beta_k^2} = \mu_{i1}\mu_{ik}(\mu_{ik} - \mu_{i1}) x_i x_i^\top.$$

For component $p = k+1$:

$$\frac{\partial^2 \mu_{i,k+1}}{\partial \beta_k^2} = \mu_{ik}(1 - \mu_{ik})(1 - 2\mu_{ik}) x_i x_i^\top.$$

For other components $p \neq 1, p \neq k+1$:

$$\frac{\partial^2 \mu_{ip}}{\partial \beta_k^2} = \mu_{ip}\mu_{ik}(\mu_{ik} + \mu_{ip})x_i x_i^\top.$$

Case 2: $k \neq k'$ (Different Components)

For component $p = 1$ (reference):

$$\frac{\partial^2 \mu_{i1}}{\partial \beta_k \partial \beta_{k'}} = \mu_{i1}\mu_{ik}\mu_{ik'} x_i x_i^\top.$$

For component $p = k+1$:

$$\frac{\partial^2 \mu_{i,k+1}}{\partial \beta_k \partial \beta_{k'}} = -\mu_{ik}\mu_{ik'}(1 - \mu_{ik}) x_i x_i^\top.$$

For component $p = k'+1$:

$$\frac{\partial^2 \mu_{i,k'+1}}{\partial \beta_k \partial \beta_{k'}} = -\mu_{ik'}\mu_{ik}(1 - \mu_{ik'}) x_i x_i^\top.$$

For other components $p \neq 1, k+1, k'+1$:

$$\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \beta_{k'}} = \mu_{ip}\mu_{ik}\mu_{ik'} x_i x_i^\top.$$

D.3.7 Assembling the Second-Order Term

The second-order correction to the Hessian is:

$$H^{(2)}_{k,k'} = \sum_{i=1}^{n}\sum_{j=1}^{d} r_{\alpha,ij} \frac{\partial^2 m_{\alpha,ij}}{\partial \beta_k \partial \beta_{k'}},$$

where

$$\frac{\partial^2 m_{\alpha,ij}}{\partial \beta_k \partial \beta_{k'}} = [H]_j \cdot \left[\frac{\partial J_u(i)}{\partial \beta_{k'}} \cdot J_\mu(i,k) + J_u(i) \cdot \frac{\partial J_\mu(i,k)}{\partial \beta_{k'}}\right] \cdot \frac{D}{\alpha}.$$

Here $[H]_j$ denotes the j-th row of the Helmert matrix.

Explicit Form

$$H^{(2)}_{k,k'} = \sum_{i=1}^{n} r_{\alpha,i}^\top \cdot \frac{D}{\alpha} H \cdot \left[\sum_{p=1}^{D}\left(\sum_{\ell,q} H_u(i,\ell,p,q)\frac{\partial \mu_{ip}}{\partial \beta_{k'}}\right) J_\mu(i,k)_q e_\ell e_q^\top + J_u(i) \cdot \frac{\partial^2 \mu}{\partial \beta_k \partial \beta_{k'}} x_i x_i^\top\right].$$

D.3.8 Complete Hessian (Exact)

$$H_{k,k'} = H^{(1)}_{k,k'} + H^{(2)}_{k,k'} = -X^\top \mathrm{diag}(W^{(1)}_{k,k'})X + \sum_{i=1}^{n} r_{\alpha,i}^\top \cdot S_{k,k'}(i) \cdot x_i x_i^\top,$$

where $W^{(1)}_{k,k'}$ are the Gauss-Newton weights and $S_{k,k'}(i)$ is the second-order correction tensor for observation i.

E Gradient vector and Hessian matrix for the α–SAR model

The α–SAR model minimizes the sum of squared errors (SSE):

$$\text{SSE}(\mathbf{Y}, \mathbf{X}; \alpha, \rho, \mathbf{B}) = \ell(\theta) = \sum_{i=1}^{n} \|\mathbf{y}_{i,\alpha} - \mu_{i,\alpha}\|^2 = \sum_{i=1}^{n} (\mathbf{y}_{i,\alpha} - \mu_{i,\alpha})^\top (\mathbf{y}_{i,\alpha} - \mu_{i,\alpha})$$

where $\theta = (\text{vec}(\mathbf{B})^\top, \rho)^\top$ contains all parameters.

E.1 Model Specification Review

The fitted compositional values are:

$$\mu_i = \begin{cases} \dfrac{1}{1 + \sum\limits_{j=1}^{d} e^{\tilde{\mathbf{x}}_i^\top \beta_j}} & \text{for } i = 1 \\ \dfrac{e^{\tilde{\mathbf{x}}_i^\top \beta_{i-1}}}{1 + \sum\limits_{j=1}^{d} e^{\tilde{\mathbf{x}}_i^\top \beta_j}} & \text{for } i = 2, \ldots, D \end{cases}$$

where:

$$\tilde{\mathbf{x}}_i = [\mathbf{S}(\rho)^{-1}\mathbf{x}]_i = [(I_n - \rho W)^{-1}\mathbf{x}]_i$$

The Hessian matrix has the block structure:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\rho} \\ \mathbf{H}_{\rho\beta} & H_{\rho\rho} \end{bmatrix},$$

where:

- $\mathbf{H}_{\beta\beta}$ is $(dp) \times (dp)$: derivatives w.r.t. regression coefficients
- $\mathbf{H}_{\beta\rho}$ is $(dp) \times 1$: mixed derivatives
- $\mathbf{H}_{\rho\beta} = \mathbf{H}_{\beta\rho}^\top$ by symmetry
- $H_{\rho\rho}$ is 1×1: second derivative w.r.t. ρ

E.2 First Derivatives (Gradient)

E.2.1 Gradient with respect to β_k

$$\frac{\partial \ell}{\partial \beta_k} = \sum_{i=1}^{n} \sum_{j=1}^{d} r_{i,\alpha,j} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k},$$

where $r_{i,\alpha,j} = y_{i,\alpha,j} - \mu_{i,\alpha,j}$ are the residuals.

Using the chain rule:

$$\frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k} = \frac{D}{\alpha} \sum_{\ell=1}^{D} \sum_{p=1}^{D} H_{j\ell} \frac{\partial u_{i\ell}}{\partial \mu_{ip}} \frac{\partial \mu_{ip}}{\partial \beta_k},$$

where $H_{j\ell}$ is the (j, ℓ) element of the Helmert sub-matrix.

E.2.2 Gradient with respect to ρ

$$\frac{\partial \ell}{\partial \rho} = \sum_{i=1}^{n} \sum_{j=1}^{d} r_{i,\alpha,j} \frac{\partial \mu_{i,\alpha,j}}{\partial \rho}.$$

Using the chain rule through $\tilde{\mathbf{x}}$:

$$\frac{\partial \mu_{i,\alpha,j}}{\partial \rho} = \sum_{p=1}^{D} \frac{\partial \mu_{i,\alpha,j}}{\partial \mu_{ip}} \frac{\partial \mu_{ip}}{\partial \rho},$$

where:

$$\frac{\partial \mu_{ip}}{\partial \rho} = \sum_{s=1}^{p} \frac{\partial \mu_{ip}}{\partial \tilde{x}_{is}} \frac{\partial \tilde{x}_{is}}{\partial \rho}.$$

E.3 Second Derivatives: $\mathbf{H}_{\beta\beta}$

The (k, k') block of the Hessian (w.r.t. β_k and $\beta_{k'}$) is:

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_{k'}^\top} = -\sum_{i=1}^{n} \sum_{j=1}^{d} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_{k'}^\top} + \sum_{i=1}^{n} \sum_{j=1}^{d} r_{i,\alpha,j} \frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \beta_{k'}^\top}.$$

E.3.1 Gauss-Newton Approximation (First Term)

The first term is the Gauss-Newton approximation:

$$\mathbf{H}_{k,k'}^{(1)} = -\sum_{i=1}^{n} \sum_{j=1}^{d} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_{k'}^\top}.$$

This can be written in matrix form:

$$\mathbf{H}_{k,k'}^{(1)} = -\tilde{\mathbf{X}}^\top \operatorname{diag}(\mathbf{W}_{k,k'}) \tilde{\mathbf{X}},$$

where $\tilde{\mathbf{X}} = \mathbf{S}(\rho)^{-1} \mathbf{X}$ and:

$$W_{k,k'}(i,i) = \sum_{j=1}^{d} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k} \cdot \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_{k'}^\top}.$$

E.3.2 Exact Correction (Second Term)

The second term involves second derivatives of the transformed composition:

$$\mathbf{H}_{k,k'}^{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{d} r_{i,\alpha,j} \frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \beta_{k'}^\top}.$$

Computing $\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \beta_{k'}^\top}$ requires:

$$\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \beta_{k'}^\top} = \frac{D}{\alpha} \sum_{\ell=1}^{D} \sum_{p=1}^{D} H_{j\ell} \left[\frac{\partial^2 u_{i\ell}}{\partial \mu_{ip} \partial \beta_{k'}^\top} \frac{\partial \mu_{ip}}{\partial \beta_k} + \frac{\partial u_{i\ell}}{\partial \mu_{ip}} \frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \beta_{k'}^\top} \right].$$

E.3.3 Second Derivative of Power Transformation

$$\frac{\partial^2 u_{i\ell}}{\partial \mu_{ip} \partial \beta_{k'}^\top} = \sum_{q=1}^{D} \frac{\partial^2 u_{i\ell}}{\partial \mu_{ip} \partial \mu_{iq}} \frac{\partial \mu_{iq}}{\partial \beta_{k'}^\top}.$$

The second derivatives of the power transformation $\frac{\partial^2 u_{i\ell}}{\partial \mu_{ip} \partial \mu_{iq}}$ are provided above.

E.3.4 Second Derivative of Multinomial Logit

For the multinomial logit part, the second derivative depends on the case:

Case 1: $k = k'$ (same component)

For component $p = 1$ (reference):

$$\frac{\partial^2 \mu_{i1}}{\partial \beta_k^2} = \mu_{i1} \mu_{ik+1} (\mu_{ik+1} - \mu_{i1}) \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

For component $p = k + 1$:

$$\frac{\partial^2 \mu_{i,k+1}}{\partial \beta_k^2} = \mu_{ik+1} (1 - \mu_{ik+1})(1 - 2\mu_{ik+1}) \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

For other components $p \neq 1, p \neq k + 1$:

$$\frac{\partial^2 \mu_{ip}}{\partial \beta_k^2} = \mu_{ip} \mu_{ik+1} (\mu_{ik+1} + \mu_{ip}) \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

Case 2: $k \neq k'$ (different components)

For component $p = 1$ (reference):

$$\frac{\partial^2 \mu_{i1}}{\partial \beta_k \partial \beta_{k'}^\top} = \mu_{i1} \mu_{ik+1} \mu_{ik'+1} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

For component $p = k + 1$:

$$\frac{\partial^2 \mu_{i,k+1}}{\partial \beta_k \partial \beta_{k'}^\top} = -\mu_{ik+1} \mu_{ik'+1} (1 - \mu_{ik+1}) \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

For component $p = k' + 1$:

$$\frac{\partial^2 \mu_{i,k'+1}}{\partial \beta_k \partial \beta_{k'}^\top} = -\mu_{ik'+1} \mu_{ik+1} (1 - \mu_{ik'+1}) \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

For other components $p \neq 1, k + 1, k' + 1$:

$$\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \beta_{k'}^\top} = \mu_{ip} \mu_{ik+1} \mu_{ik'+1} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top.$$

E.3.5 Mixed Derivatives: $\mathbf{H}_{\beta\rho}$

The mixed derivative block is:

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \rho} = -\sum_{i=1}^{n} \sum_{j=1}^{d} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k} \frac{\partial \mu_{i,\alpha,j}}{\partial \rho} + \sum_{i=1}^{n} \sum_{j=1}^{d} r_{i,\alpha,j} \frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \rho}.$$

E.3.6 Gauss-Newton Part

$$\mathbf{H}^{(1)}_{\beta_k,\rho} = -\sum_{i=1}^{n}\sum_{j=1}^{d} \frac{\partial \mu_{i,\alpha,j}}{\partial \beta_k}\frac{\partial \mu_{i,\alpha,j}}{\partial \rho}.$$

E.3.7 Exact Correction

$$\mathbf{H}^{(2)}_{\beta_k,\rho} = \sum_{i=1}^{n}\sum_{j=1}^{d} r_{i,\alpha,j}\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \rho}.$$

The mixed derivative requires:

$$\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \beta_k \partial \rho} = \frac{D}{\alpha}\sum_{\ell=1}^{D}\sum_{p=1}^{D} H_{j\ell}\left[\frac{\partial^2 u_{i\ell}}{\partial \mu_{ip}\partial \rho}\frac{\partial \mu_{ip}}{\partial \beta_k} + \frac{\partial u_{i\ell}}{\partial \mu_{ip}}\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \rho}\right].$$

E.3.8 Computing $\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \rho}$

Using the chain rule:

$$\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \rho} = \sum_{q=1}^{D} \frac{\partial^2 \mu_{ip}}{\partial \mu_{iq}\partial \beta_k}\frac{\partial \mu_{iq}}{\partial \rho} + \frac{\partial \mu_{ip}}{\partial \beta_k}\frac{\partial \tilde{\mathbf{x}}_i}{\partial \rho}.$$

The key term is:

$$\frac{\partial}{\partial \rho}\left[\frac{\partial \mu_{ip}}{\partial \beta_k}\right] = \frac{\partial \mu_{ip}}{\partial \beta_k \partial \tilde{\mathbf{x}}_i}\frac{\partial \tilde{\mathbf{x}}_i}{\partial \rho}.$$

Since $\frac{\partial \tilde{\mathbf{x}}_i}{\partial \rho} = \mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i$, we have the following expressions.

For component $p = 1$:

$$\frac{\partial^2 \mu_{i1}}{\partial \beta_k \partial \rho} = -\mu_{i1}\mu_{ik+1}[\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i] + \mu_{i1}\mu_{ik+1}(2\mu_{i1}-1)\tilde{\mathbf{x}}_i \cdot [\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i]^\top \beta_k.$$

For component $p = k+1$:

$$\frac{\partial^2 \mu_{i,k+1}}{\partial \beta_k \partial \rho} = \mu_{ik+1}(1-\mu_{ik+1})[\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i] + \mu_{ik+1}(1-\mu_{ik+1})(1-2\mu_{ik+1})\tilde{\mathbf{x}}_i \cdot [\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i]^\top \beta_k.$$

For other components $p \neq 1, k+1$:

$$\frac{\partial^2 \mu_{ip}}{\partial \beta_k \partial \rho} = -\mu_{ip}\mu_{ik+1}[\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i] + \mu_{ip}\mu_{ik+1}(\mu_{ip}+\mu_{ik+1})\tilde{\mathbf{x}}_i \cdot [\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i]^\top \beta_k.$$

E.3.9 Second Derivative: $H_{\rho\rho}$

The pure second derivative with respect to ρ is:

$$\frac{\partial^2 \ell}{\partial \rho^2} = -\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\frac{\partial \mu_{i,\alpha,j}}{\partial \rho}\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{d} r_{i,\alpha,j}\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \rho^2}.$$

E.3.10 Gauss-Newton Part

$$H^{(1)}_{\rho\rho} = -\sum_{i=1}^{n}\sum_{j=1}^{d}\left(\frac{\partial \mu_{i,\alpha,j}}{\partial \rho}\right)^2.$$

E.3.11 Exact Correction

$$H_{\rho\rho}^{(2)} = \sum_{i=1}^{n}\sum_{j=1}^{d} r_{i,\alpha,j}\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \rho^2}.$$

The second derivative requires:

$$\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \rho^2} = \sum_{p=1}^{D}\frac{\partial^2 \mu_{i,\alpha,j}}{\partial \mu_{ip}^2}\left(\frac{\partial \mu_{ip}}{\partial \rho}\right)^2 + \sum_{p=1}^{D}\frac{\partial \mu_{i,\alpha,j}}{\partial \mu_{ip}}\frac{\partial^2 \mu_{ip}}{\partial \rho^2}.$$

E.3.12 Computing $\frac{\partial^2 \mu_{ip}}{\partial \rho^2}$

This requires the second derivative of $\tilde{\mathbf{x}}_i$ with respect to ρ:

$$\begin{aligned}
\frac{\partial^2 \tilde{\mathbf{x}}_i}{\partial \rho^2} &= \frac{\partial}{\partial \rho}[\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i] = \mathbf{S}(\rho)^{-1}W\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i + \mathbf{S}(\rho)^{-1}W\frac{\partial \tilde{\mathbf{x}}_i}{\partial \rho} \\
\frac{\partial^2 \tilde{\mathbf{x}}_i}{\partial \rho^2} &= 2\mathbf{S}(\rho)^{-1}W\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i.
\end{aligned}$$

Then:

$$\frac{\partial^2 \mu_{ip}}{\partial \rho^2} = \sum_{q=1}^{D}\frac{\partial^2 \mu_{ip}}{\partial \mu_{iq}^2}\left(\frac{\partial \mu_{iq}}{\partial \rho}\right)^2 + \sum_{q=1}^{D}\frac{\partial \mu_{ip}}{\partial \mu_{iq}}\frac{\partial^2 \mu_{iq}}{\partial \rho^2} + \sum_{s=1}^{p}\frac{\partial \mu_{ip}}{\partial \tilde{x}_{is}}\frac{\partial^2 \tilde{x}_{is}}{\partial \rho^2}.$$

For component $p = 1$:

$$\frac{\partial^2 \mu_{i1}}{\partial \rho^2} = \mu_{i1}\sum_{k=1}^{d}\mu_{ik+1}\left[(\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i)^\top\beta_k\right]^2(2\mu_{i1}-1) - \mu_{i1}\sum_{k=1}^{d}\mu_{ik+1}\cdot 2[\mathbf{S}(\rho)^{-1}W\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i]^\top\beta_k.$$

For component $p = k+1$:

$$\frac{\partial^2 \mu_{ik+1}}{\partial \rho^2} = \mu_{ik+1}(1-\mu_{ik+1})[(\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i)^\top\beta_k]^2(1-2\mu_{ik+1}) + \mu_{ik+1}(1-\mu_{ik+1})\cdot 2[\mathbf{S}(\rho)^{-1}W\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i]^\top\beta_k.$$

For other components $p \neq 1, k+1$:

$$\frac{\partial^2 \mu_{ip}}{\partial \rho^2} = \mu_{ip}\sum_{k=1}^{d}\mu_{ik+1}[(\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i)^\top\beta_k]^2(\mu_{ip}+\mu_{ik+1}) - \mu_{ip}\sum_{k=1}^{d}\mu_{ik+1}\cdot 2[\mathbf{S}(\rho)^{-1}W\mathbf{S}(\rho)^{-1}W\tilde{\mathbf{x}}_i]^\top\beta_k.$$

E.3.13 Complete Hessian Matrix

The complete Hessian is:

$$\mathbf{H} = \mathbf{H}^{(1)} + \mathbf{H}^{(2)},$$

where $\mathbf{H}^{(1)}$ is the Gauss-Newton approximation (always negative semi-definite) and $\mathbf{H}^{(2)}$ is the exact correction involving second derivatives weighted by residuals.

E.3.14 Simplified Structure

In matrix-vector notation:

$$\mathbf{H} = \begin{bmatrix} -\tilde{\mathbf{X}}^\top\mathbf{W}_{\beta\beta}\tilde{\mathbf{X}} + \mathbf{R}_{\beta\beta} & \mathbf{h}_{\beta\rho} \\ \mathbf{h}_{\beta\rho}^\top & h_{\rho\rho} \end{bmatrix},$$

where:

- $\mathbf{W}_{\beta\beta}$ contains the Gauss-Newton weights
- $\mathbf{R}_{\beta\beta}$ contains the residual-weighted second derivatives
- $\mathbf{h}_{\beta\rho}$ contains the mixed derivatives
- $h_{\rho\rho}$ is the second derivative w.r.t. ρ

F Standard error of the SMEs of the α–SLX model

The standard errors of the SMEs are computed again via the delta method, similarly to (10). Thus we need to compute the Jacobians. For observation i, component ℓ, and covariate k we can define the following cases:

Jacobian for the direct SMEs

$$\frac{\partial DE_{i\ell k}}{\partial \beta_{ms}} = \begin{cases} -\delta_{sk}\mu_{i1}\mu_{im+1} - \frac{\partial \mu_{i1}}{\partial \beta_{ms}} \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1} - \mu_{i1} \sum_{j=1}^{d} \beta_{jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{ms}} & \text{for } \ell = 1 \\ \delta_{sk}\delta_{m,\ell-1}\mu_{i\ell} - \delta_{sk}\mu_{i\ell}\mu_{im+1} + \frac{\partial \mu_{i\ell}}{\partial \beta_{ms}} \left(\beta_{\ell-1,k} - \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1} \right) & \\ -\mu_{i\ell} \sum_{j=1}^{d} \beta_{jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{ms}} & \text{for } \ell = 2, \ldots, D. \end{cases}$$

Jacobian for the indirect SMEs

$$\frac{\partial IDE_{i\ell k}}{\partial \gamma_{ms}} = \begin{cases} -\delta_{sk}\mu_{i1}\mu_{im+1} - \frac{\partial \mu_{i1}}{\partial \gamma_{ms}} \sum_{j=1}^{d} \gamma_{jk}\mu_{ij+1} - \mu_{i1} \sum_{j=1}^{d} \gamma_{jk} \frac{\partial \mu_{ij+1}}{\partial \gamma_{ms}} & \text{for } \ell = 1 \\ \delta_{sk}\delta_{m,\ell-1}\mu_{i\ell} - \delta_{sk}\mu_{i\ell}\mu_{im+1} + \frac{\partial \mu_{i\ell}}{\partial \gamma_{ms}} \left(\gamma_{\ell-1,k} - \sum_{j=1}^{d} \gamma_{jk}\mu_{ij+1} \right) & \\ -\mu_{i\ell} \sum_{j=1}^{d} \gamma_{jk} \frac{\partial \mu_{ij+1}}{\partial \gamma_{ms}} & \text{for } \ell = 2, \ldots, D. \end{cases}$$

For both cases, the necessary derivatives of μ with respect to each β and γ coefficient are given by

$$\frac{\partial \mu_{ip}}{\partial \beta_{rs}} = \begin{cases} -\mu_{i1}\mu_{ir+1}x_{is} & \text{if } p = 1 \\ \mu_{ip}(1-\mu_{ip})x_{is} & \text{if } p = r+1 \\ -\mu_{ip}\mu_{ir+1}x_{is} & \text{otherwise} \end{cases} \quad \text{and} \quad \frac{\partial \mu_{ip}}{\partial \gamma_{rs}} = \begin{cases} -\mu_{i1}\mu_{ir+1}(\mathbf{W}\mathbf{x})_{is} & \text{if } p = 1 \\ \mu_{ip}(1-\mu_{ip})(\mathbf{W}\mathbf{x})_{is} & \text{if } p = r+1 \\ -\mu_{ip}\mu_{ir+1}(\mathbf{W}\mathbf{x})_{is} & \text{otherwise.} \end{cases}$$

The Jacobian for the total SMEs is simply the sum of the Jacobians for the direct and indirect SMEs. The covariance matrices for each type of SME is the same as in the case of the α–regression, Eqs. (10) and (11).

G Standard error of the SMEs of the α–SAR model

For the SMEs, we need to account for uncertainty in the parameters $\theta = (\mathrm{vec}(\mathbf{B})^\top, \rho)^\top$, where $\mathbf{B} = (\beta_1, \ldots, \beta_d)$ is the $d \times p$ matrix of coefficients. The complete Jacobian for observation i, component ℓ, covariate k has dimension $1 \times (dp+1)$:

$$\mathbf{J}_{i\ell k} = \left[\frac{\partial ME_{i\ell k}}{\partial \beta_{11}}, \ldots, \frac{\partial ME_{i\ell k}}{\partial \beta_{dp}}, \frac{\partial ME_{i\ell k}}{\partial \rho}\right]$$

The effective covariate is

$$\tilde{x} = S(\rho)^{-1}x = (I_n - \rho W)^{-1}x.$$

Jacobian of the direct SMEs

The derivative of the direct SME is

$$\frac{\partial DE_{i\ell k}}{\partial \beta_{rs}} = \frac{\partial^2 \mu_{i\ell}}{\partial \tilde{x}_k \partial \beta_{rs}} \cdot [S(\rho)^{-1}]_{ii}$$

This requires the second derivative of the composition with respect to the transformed covariate and the regression coefficients. From the multinomial logit structure:

$$\frac{\partial \mu_{ip}}{\partial \beta_{rs}} = \begin{cases} -\mu_{i1}\mu_{ir+1}\tilde{x}_{is} & \text{if } p = 1 \\ \mu_{ip}(1-\mu_{ip})\tilde{x}_{is} & \text{if } p = r+1 \\ -\mu_{ip}\mu_{ir+1}\tilde{x}_{is} & \text{otherwise,} \end{cases}$$

where $\tilde{x}_{is}$ is the s-th component of $\tilde{x}_i = S(\rho)^{-1}x_i$.

For $\ell = 1$ (reference category):

$$\frac{\partial^2 \mu_{i1}}{\partial \tilde{x}_k \partial \beta_{rs}} = -\frac{\partial \mu_{i1}}{\partial \beta_{rs}} \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1} - \mu_{i1} \sum_{j=1}^{d} \beta_{jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{rs}} - \mu_{i1}\delta_{rk}\mu_{ir+1},$$

where δ_{rk} is the Kronecker delta ($\delta_{rk} = 1$ if $r = k$, 0 otherwise). For $\ell = 2, \ldots, D$:

$$\frac{\partial^2 \mu_{i\ell}}{\partial \tilde{x}_k \partial \beta_{rs}} = \frac{\partial \mu_{i\ell}}{\partial \beta_{rs}} \left[\beta_{\ell-1,k} - \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1}\right] + \mu_{i\ell} \left[\delta_{\ell-1,r}\delta_{sk} - \delta_{rk}\mu_{ir+1} - \sum_{j=1}^{d} \beta_{jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{rs}}\right]$$

The complete formula for $\frac{\partial DE_{i\ell k}}{\partial \beta_{rs}}$ is

For $\ell = 1$:

$$\frac{\partial DE_{i\ell k}}{\partial \beta_{rs}} = [S(\rho)^{-1}]_{ii} \left[-\frac{\partial \mu_{i1}}{\partial \beta_{rs}} \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1} - \mu_{i1} \sum_{j=1}^{d} \beta_{jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{rs}} - \mu_{i1}\delta_{rk}\mu_{ir+1}\right]$$

For $\ell = 2, \ldots, D$:

$$\frac{\partial DE_{i\ell k}}{\partial \beta_{rs}} = [S(\rho)^{-1}]_{ii} \left[\frac{\partial \mu_{i\ell}}{\partial \beta_{rs}} \left[\beta_{\ell-1,k} - \sum_{j=1}^{d} \beta_{jk}\mu_{ij+1}\right] + \mu_{i\ell} \left[\delta_{\ell-1,r}\delta_{sk} - \delta_{rk}\mu_{ir+1} - \sum_{j=1}^{d} \beta_{jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{rs}}\right]\right]$$

The derivative of the direct SMEs with respect to ρ involves both the change in $\mu_{i\ell}$ through $\tilde{x}$ and the change in $[S(\rho)^{-1}]_{ii}$:

$$\frac{\partial DE_{i\ell k}}{\partial \rho} = \frac{\partial}{\partial \rho}\left[\frac{\partial \mu_{i\ell}}{\partial \tilde{x}_k}\right] \cdot [S(\rho)^{-1}]_{ii} + \frac{\partial \mu_{i\ell}}{\partial \tilde{x}_k} \cdot \frac{\partial [S(\rho)^{-1}]_{ii}}{\partial \rho},$$

where

$$\frac{\partial [S(\rho)^{-1}]_{ii}}{\partial \rho} = [S(\rho)^{-1} W S(\rho)^{-1}]_{ii}$$

The derivative of $\tilde{x}$ with respect to ρ is

$$\frac{\partial \tilde{x}}{\partial \rho} = \frac{\partial S(\rho)^{-1}}{\partial \rho} x = S(\rho)^{-1} W S(\rho)^{-1} x = S(\rho)^{-1} W \tilde{x}$$

and via the chain rule application we obtain

$$\frac{\partial}{\partial \rho}\left[\frac{\partial \mu_{i\ell}}{\partial \tilde{x}_k}\right] = \sum_{p=1}^{D} \frac{\partial^2 \mu_{i\ell}}{\partial \tilde{x}_k \partial \mu_{ip}} \cdot \frac{\partial \mu_{ip}}{\partial \rho},$$

where

$$\frac{\partial \mu_{ip}}{\partial \rho} = \sum_{s=1}^{p} \frac{\partial \mu_{ip}}{\partial \tilde{x}_s} \cdot [S(\rho)^{-1} W \tilde{x}]_{is}$$

Finally, we get

$$\frac{\partial DE_{i\ell k}}{\partial \rho} = \left[\sum_{p=1}^{D} \frac{\partial^2 \mu_{i\ell}}{\partial \tilde{x}_k \partial \mu_{ip}} \cdot \sum_{s=1}^{p} \frac{\partial \mu_{ip}}{\partial \tilde{x}_s} \cdot [S(\rho)^{-1} W \tilde{x}]_{is}\right] \cdot [S(\rho)^{-1}]_{ii} + \frac{\partial \mu_{i\ell}}{\partial \tilde{x}_k} \cdot [S(\rho)^{-1} W S(\rho)^{-1}]_{ii}.$$

Jacobian for the indirect SMEs

The derivative with respect to β_{rs} is

$$\frac{\partial IE_{i\ell k}}{\partial \beta_{rs}} = \frac{\partial^2 \mu_{i\ell}}{\partial \tilde{x}_k \partial \beta_{rs}} \cdot \sum_{j \neq i} [S(\rho)^{-1}]_{ij}.$$

This uses the same $\frac{\partial^2 \mu_{i\ell}}{\partial \tilde{x}_k \partial \beta_{rs}}$ as for direct effects.

The derivative with respect to ρ is

$$\frac{\partial IE_{i\ell k}}{\partial \rho} = \frac{\partial}{\partial \rho}\left[\frac{\partial \mu_{i\ell}}{\partial \tilde{x}_k}\right] \cdot \sum_{j \neq i} [S(\rho)^{-1}]_{ij} + \frac{\partial \mu_{i\ell}}{\partial \tilde{x}_k} \cdot \sum_{j \neq i} \frac{\partial [S(\rho)^{-1}]_{ij}}{\partial \rho},$$

where

$$\sum_{j \neq i} \frac{\partial [S(\rho)^{-1}]_{ij}}{\partial \rho} = \sum_{j \neq i} [S(\rho)^{-1} W S(\rho)^{-1}]_{ij}.$$

Jacobian vector for the SMEs of the GW$_\alpha$R model at location (ν_i, u_i)

Define $\theta_i = \mathrm{vec}(\mathbf{B}_i) \in \mathbb{R}^{dp}$, where $\mathbf{B}_i = (\beta_{i,1}, \ldots, \beta_{i,d})$ is the $d \times p$ matrix the of location-specific coefficients (at (ν_i, u_i)). For observation i, component ℓ, and covariate k, at location (ν_i, u_i) the MEs are defined as

$$\frac{\partial \mu_{i\ell k}}{\partial \beta_{i,ms}} = \begin{cases} -\delta_{sk}\mu_{i1}\mu_{im+1} - \frac{\partial \mu_{i1}}{\partial \beta_{i,ms}} \sum_{j=1}^{d} \beta_{i,jk}\mu_{ij+1} - \mu_{i1} \sum_{j=1}^{d} \beta_{i,jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{i,ms}} & \text{for } \ell = 1 \\ \delta_{sk}\delta_{m,\ell-1}\mu_{i\ell} - \delta_{sk}\mu_{i\ell}\mu_{im+1} + \frac{\partial \mu_{i\ell}}{\partial \beta_{i,ms}}\left(\beta_{i,\ell-1,k} - \sum_{j=1}^{d} \beta_{i,jk}\mu_{ij+1}\right) & \\ -\mu_{i\ell} \sum_{j=1}^{d} \beta_{i,jk} \frac{\partial \mu_{ij+1}}{\partial \beta_{i,ms}} & \text{for } \ell = 2, \ldots, D, \end{cases}$$

where the derivatives of the composition with respect to the location-specific coefficients are:

$$\frac{\partial \mu_{ip}}{\partial \beta_{i,rs}} = \begin{cases} -\mu_{i1}\mu_{ir+1}x_{is} & \text{if } p = 1 \\ \mu_{ip}(1-\mu_{ip})x_{is} & \text{if } p = r+1 \\ -\mu_{ip}\mu_{ir+1}x_{is} & \text{otherwise,} \end{cases}$$

where $r \in \{1, \ldots, d\}$ and $s \in \{1, \ldots, p\}$.