Quasi-Maximum Likelihood Estimation for Spatial Panel Data Regressions

Zhenlin Yang
School of Economics and Social Sciences
Singapore Management University, Singapore 259756
zlyang@smu.edu.sg

First Version, May 2005; This version, July 2006

Abstract

This article considers quasi-maximum likelihood estimations (QMLE) for two spatial panel data regression models: mixed effects model with spatial errors and transformed mixed effects model (where response and covariates are transformed) with spatial errors. One aim of transformation is to normalize the data, thus the transformed models are more robust with respect to the normality assumption compared with the standard ones. QMLE method provides additional protection against violation of normality assumption. Asymptotic properties of the QMLEs are investigated. Numerical illustrations are provided.

JEL Classification: C21, C23, C51

Keywords: Asymptotics; Flexible functional form; Fixed effects; Quasi-maximum likelihood; Random Effects; Spatial error correlation; Demand equation.

1 Introduction

In recent number of years, spatial panel data regression has received an increasing attention by the researchers. See, for example, Baltagi, Song and Koh (2003), Baltagi and Li (2004), Baltagi, Song and Jung (2004), Persaran (2002, 2004), Azomahou (1999, 2000, and 2001), and Elhorst (2003). For very recent developments, see papers presented at the *Spatial Econometrics Workshop, Kiel, Germany 2005* available at `http://www.uni-kiel.de/ifw/konfer/spatial/prel-program.htm`, and papers presented at *International Workshop on Spatial Econometrics and Statistics, Rome, Italy 2006*. However, important issues and techniques such as quasi-maximum likelihood estimation and data transformation, in particular the latter, have not been considered in the spatial panel framework.

In this article, we consider quasi-maximum likelihood estimations (QMLE) for two spatial panel data models: (i) mixed effects model with spatial errors; and (ii) transformed mixed effects model with spatial errors. By mixed effects model we mean a panel model with fixed time effects and random individual effects. Transformation can be applied to both the response and some of the covariates. Model (i) is a standard one, but QMLE has not been considered. The transformed model is an extension of the standard model.

In a recent paper, Lee (2004) considered the asymptotic distributions of the QMLEs of a cross-sectional regression model with spatial lag. His work largely motivated our work in a panel set up. QMLE method provides robust standard errors: robust against misspecification on error distributions. Transformation aims to bring the data to near normality, induce flexible functional form, induce simpler model structure and reduce heteroscedasticity (Box and Cox, 1964). Thus, transformation together with QMLE method offer two-way protections against nonnormality of data. These features make the transformed spatial panel models very attractive.

Research in spatial panel data regression often assumes that the data (in original or log form) follow normal distributions; see, for example, Anselin (1988, Sec. 10.2);

Baltagi and Li (2004); Baltagi, Song and Koh (2003); Baltagi, Song and Jung (2004). However, in practical applications, economic data are often non-normal, hence it is necessary to transform the data before fitting the model.[1] While this model is preferable for modelling economic panel data, it renders the standard estimation techniques such as the generalized least squares (GLS) and generalized method of moments (GMM) unapplicable (Davidson and MacKinnon, 1993, p. 243). In this sense, the QMLE method comes in as a natural choice for estimating the proposed transformation model. A practical issue with the use of the maximum likelihood estimation (MLE) method or QMLE method is its computational complexity. We show in this paper that the amount of computation involved is feasible for a desktop computer for data sets of moderate sizes.

Most of the economic panel data are of the feature that there many cross sections and each cross section corresponds to a short time period. This makes it theoretically possible and practically popular to consider the unobservable time effects as fixed. Furthermore, various policy interventions over time also justify the use of time dummies for controlling the unobservable time effects (Baltagi, et al., 2000; Hamilton, 1972; Baltagi and Levin, 1986). In contrast, for the time-invariant, individual-specific effects, incidental parameters problem prevents the consideration of fixed individual effects. However, this problem can be resolved by including individual-specific variables and treating the 'left-over' unobservable individual effects as random.

In applying our models and methods to the state demand for cigarette data, we found (i) strong evidence for the existence of spatial effect, (ii) strong evidence for the use of general Box-Cox functional form rather than the traditional log-log form, (iii) strong evidence for the existence of fixed time effects as well as random state effects. These results have implications for more accurate prediction in cigarette sales

[1]Several authors, including van Gastel and Paelinck (1995), Griffith et al. (1998), Baltagi and Li (2001), and Pace, et al. (2001), have discussed and demonstrated the importance of transformations in analyzing spatial effects based on cross-sectional data.

compared with Baltagi and Li (2004). We also found that the QMLE standard errors are larger than the MLE standard errors for most of the parameter estimates. As a result, the corresponding t-ratios are smaller.

The rest of paper is organized as follows. Section 2 presents the two models and develops the MLE or QMLE procedure for model estimation. Section 3 considers asymptotic properties of the QMLEs, including the consistency and asymptotic normality. Section 4 presents an empirical illustration using the states demand for cigarettes data. Section 5 concludes the paper.

2 Quasi-Maximum Likelihood Estimation

We now develop the QMLE procedures for estimating the models. Some of the matrix differential formulas that are useful in our derivation can be found in Magnus (1982), Griffith (1981), or Magnus and Neudecker (1999).

2.1 Mixed effects model with spatial errors

The mixed effects model with spatial errors has the form

$$Y_{ti} = X_{ti}'\beta_1 + \eta_t + \mu_i + \epsilon_{ti},\ t = 1, \cdots, T; i = 1, \cdots, N,$$

where $\{\eta_t\}$ are the fixed time effects, $\{\mu_i\}$ are the random individual effects, and $\{\epsilon_{it}\}$ are the spatially correlated errors. In vector form

$$Y_t = X_t\beta_1 + \eta_t 1_N + \mu + \epsilon_t,\ \text{with } \epsilon_t = \delta W\epsilon_t + v_t,\ t = 1, \cdots, T, \tag{1}$$

where $Y_t = (Y_{t1}, \cdots, Y_{tN})'$, $\epsilon_t = (\epsilon_{t1}, \cdots, \epsilon_{tN})'$, $\mu = (\mu_1, \cdots, \mu_N)'$, 1_N is an N-vector of ones, and X_t is a matrix whose ith row contains the values of covariates corresponding to the ith spatial unit. The $\{\mu_i\}$ are independent and identically distributed (iid) random variables with mean 0 and variance σ_μ^2, $\{v_{ti}\}$ are iid random variables with mean 0 and variance σ_v^2, and μ_i is independent of ϵ_{ti} for all t and i. The parameter

η_t represents the fixed effect for tth time period. The parameter δ is the spatial autoregressive coefficient with $|\delta| < 1$, W is a known $N \times N$ spatial weight matrix whose diagonal elements are zero, which satisfies the condition such that $(I_N - \delta W)$ is non-singular for all $|\delta| < 1$.

Let $B = I_N - \delta W$ with I_N being an $N \times N$ identity matrix. We have $\epsilon_t = B^{-1} v_t$. The model can be rewritten conveniently in matrix notation as

$$Y = X\beta + u, \text{ with } u = (1_T \otimes I_N)\mu + (I_T \otimes B^{-1})v, \tag{2}$$

where $\otimes$ denotes the Kronecker product, $Y = (Y_1', \cdots, Y_T')'$, a $TN \times 1$ vector of responses, and $X = \{(X_1', \cdots, X_T')', (I_T \otimes 1_N)\}$, a $TN \times k$ matrix whose rows contain the values of the covariates and the dummy variables associated with the fixed effects $\{\eta_t\}$,[2] and $\beta = (\beta_1', \eta_1, \cdots, \eta_T)'$. Letting $\phi = \sigma_\mu^2/\sigma_v^2$, we have $\text{Cov}(u) = \sigma_v^2 \Omega$ with

$$\Omega = \phi(J_T \otimes I_N) + I_T \otimes (B'B)^{-1}.$$

Denoting $u = Y - X\beta$, the quasi-log likelihood function under the assumption that errors are normally distributed has the form, besides an additive constant,

$$\ell(\beta, \sigma_v^2, \phi, \delta) = -\frac{TN}{2}\log(\sigma_v^2) - \frac{1}{2}\log|\Omega| - \frac{1}{2\sigma_v^2} u'\Omega^{-1}u. \tag{3}$$

Maximization of (3) gives the MLEs if the errors are truly normal, otherwise the QMLEs. It should be noted that when large panel data are involved, the above maximization process can be quite involved computationally. Following procedures lighten the computational burden considerably and those procedures can also be generalized to the more complicated transformation model to be considered in the next subsection. First, the dimension of maximization can be reduced by concentrating out β and σ_v^2 from $\ell(\beta, \sigma_v^2, \phi, \delta)$. Given $\theta = (\phi, \delta)'$, ℓ is maximized at

$$\begin{aligned} \hat{\beta}(\theta) &= [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}Y, \\ \hat{\sigma}_v^2(\theta) &= \frac{1}{NT}\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta), \end{aligned}$$

[2]In cases that the model includes an intercept, the number of time dummies has to be reduced by one, or one constraint be put on these dummies, to ensure parameters identifiability.

where $\hat{u}(\theta) = Y - X\hat{\beta}(\theta)$. Substituting $\hat{\beta}(\theta)$ and $\hat{\sigma}_v^2(\theta)$ into the quasi-log likelihood function (3) for β and σ_v^2, respectively, gives the concentrated quasi-log likelihood after dropping the constant,

$$\ell_{\max}(\theta) = -\frac{TN}{2}\log[\hat{\sigma}_v^2(\theta)] - \frac{1}{2}\log|\Omega|. \tag{4}$$

Maximizing $\ell_{\max}(\theta)$, subject to $|\delta| < 1$, gives the QMLE $\hat{\theta}$, which upon substitution gives the unconstrained QMLEs $\hat{\beta} = \hat{\beta}(\hat{\theta})$ and $\hat{\sigma}_v^2 = \hat{\sigma}_v^2(\hat{\theta})$ for β and σ_v^2, respectively. Further, the unconstrained QMLE of σ_μ^2 is given by $\hat{\sigma}_\mu^2 = \hat{\phi}\hat{\sigma}_v^2$.

Maximization of $\ell_{\max}(\theta)$ can be facilitated by providing the analytical gradients or concentrated quasi-scores, which can be obtained by either directly differentiating $\ell_{\max}(\theta)$ with respect to ϕ and δ, or substituting $\hat{\beta}(\theta)$ and $\hat{\sigma}_v^2(\theta)$ into the last two elements of the gradient vector (see Appendix A):

$$G_\phi(\theta) = \frac{1}{2}\left(\frac{NT\hat{u}'(\theta)\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}\hat{u}(\theta)}{\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta)} - \mathrm{tr}[\Omega^{-1}(J_T \otimes I_N)]\right) \tag{5}$$

$$G_\delta(\theta) = \frac{1}{2}\left(\frac{TN\hat{u}'(\theta)\Omega^{-1}(I_T \otimes A)\Omega^{-1}\hat{u}(\theta)}{\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta)} - \mathrm{tr}[\Omega^{-1}(I_T \otimes A)]\right) \tag{6}$$

where $A = (\partial/\partial\delta)(B'B)^{-1} = (B'B)^{-1}(W'B + B'W)(B'B)^{-1}$.

The above maximization process involves repeated evaluations of Ω^{-1} and $|\Omega|$ for the $TN \times TN$ matrix Ω, which can be a great burden computationally when N or T or both are large. Following Magnus (1982), the calculations involving the $TN \times TN$ matrix Ω can be reduced to the calculations involving the $N \times N$ matrix B:

$$|\Omega| = |(B'B)^{-1} + \phi T I_N| \cdot |B|^{-2(T-1)}, \tag{7}$$

$$\Omega^{-1} = (1/T)J_T \otimes [(B'B)^{-1} + \phi T I_N]^{-1} + [I_T - (1/T)J_T] \otimes (B'B). \tag{8}$$

Following Griffith (1988, Table 3.1), calculation of $|\Omega|$ can be further simplified as

$$|B| = \prod_{i=1}^{N}(1 - \delta w_i), \text{ and } |(B'B)^{-1} + \phi T I_N| = \prod_{i=1}^{N}[(1 - \delta w_i)^{-2} + T\phi], \tag{9}$$

where w_i are the eigenvalues of W. Those formulas simplify the computations greatly and make the model estimation involving a large panel data possible.

2.2 Transformed mixed effects model with spatial errors

The transformed mixed effects model with spatial errors is

$$h(Y_t, \lambda) = Z_t\beta_1 + h(X_t, \lambda)\beta_2 + \eta_t 1_N + \mu + \epsilon_t, \text{ with } \epsilon_t = \delta W\epsilon_t + v_t, \tag{10}$$

where $h(\cdot, \lambda)$ is a monotonic transformation, known except the indexing parameter λ, called the *transformation parameter*. The Z_t matrix contains the column of ones, values of dummy variables, as well as values for those variables that do not need to be transformed. The X_t matrix contains values for those variables that are of the continuous type similar in nature to the response, and hence also need to be transformed. The error specifications are the same as in Model (1). In matrix notation, the model takes the form

$$h(Y, \lambda) = X(\lambda)\beta + u, \text{ with } u = (1_T \otimes I_N)\mu + (I_T \otimes B^{-1})v \tag{11}$$

where $h(Y, \lambda)$ is a $TN \times 1$ vector of transformed responses, and $X(\lambda)$ is a $TN \times k$ matrix whose rows contain the (transformed) values of the covariates including the time dummies. The other quantities are defined similarly to those in Section 2.1. The quasi-log likelihood function (assuming the errors are normal) has the form, besides an additive constant,

$$\ell(\beta, \sigma_v^2, \theta) = -\frac{TN}{2}\log(\sigma_v^2) - \frac{1}{2}\log|\Omega| - \frac{1}{2\sigma_v^2}u'\Omega^{-1}u + J(\lambda), \tag{12}$$

where $\theta = (\phi, \delta, \lambda)'$, $u = h(Y, \lambda) - X(\lambda)\beta$, and $J(\lambda) = \sum_{t=1}^{T}\sum_{i=1}^{N}\log h_Y(Y_{ti}, \lambda)$.

Maximization of (12) results in MLEs if the errors are exactly normal, otherwise QMLEs for the model parameters. Clearly, the addition of transformation in the model makes parameter estimation more challenging. Direct maximization of (12) may be impractical and method of simplification should be sought after. Firstly, the dimension of maximization can be reduced by concentrating out the parameters β and σ_v^2 from $\ell(\beta, \sigma_v^2, \theta)$. For given θ, ℓ is maximized at

$$\begin{aligned}
\hat{\beta}(\theta) &= [X'(\lambda)\Omega^{-1}X(\lambda)]^{-1}X'(\lambda)\Omega^{-1}h(Y, \lambda), \\
\hat{\sigma}_v^2(\theta) &= \frac{1}{NT}\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta),
\end{aligned}$$

where $\hat{u}(\theta) = h(Y,\lambda) - X(\lambda)\hat{\beta}(\phi,\delta,\lambda)$. Substituting $\hat{\beta}(\theta)$ and $\hat{\sigma}_v^2(\theta)$ into the quasi-log likelihood function (12) for β and σ_v^2, respectively, gives the concentrated quasi-log likelihood after dropping the constant,

$$\ell_{\max}(\theta) = -\frac{TN}{2}\log[\hat{\sigma}_v^2(\theta)] - \frac{1}{2}\log|\Omega| + J(\lambda). \tag{13}$$

Maximizing $\ell_{\max}(\theta)$, subject to $|\delta| < 1$, gives the QMLE $\hat{\theta}$, which upon substitution gives the unconstrained QMLEs $\hat{\beta} = \hat{\beta}(\hat{\theta})$ and $\hat{\sigma}_v^2 = \hat{\sigma}_v^2(\hat{\theta})$ for β and σ_v^2, respectively. Further, the unconstrained QMLE of σ_μ^2 is given by $\hat{\sigma}_\mu^2 = \hat{\phi}\hat{\sigma}_v^2$.

Secondly, maximization of $\ell_{\max}$ can be facilitated by providing the analytical gradients or concentrated quasi-scores, which can be obtained by either differentiating $\ell_{\max}(\theta)$ with respect to, ϕ, δ and λ, respectively, or substituting $\hat{\beta}(\theta)$ and $\hat{\sigma}_v^2(\theta)$ into the last three elements of the full gradient vector (see Appendix A):

$$G_\phi(\theta) = \frac{1}{2}\left(\frac{NT\hat{u}'(\theta)\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}\hat{u}(\theta)}{\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta)} - \mathrm{tr}[\Omega^{-1}(J_T \otimes I_N)]\right) \tag{14}$$

$$G_\delta(\theta) = \frac{1}{2}\left(\frac{TN\hat{u}'(\theta)\Omega^{-1}(I_T \otimes A)\Omega^{-1}\hat{u}(\theta)}{\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta)} - \mathrm{tr}[\Omega^{-1}(I_T \otimes A)]\right) \tag{15}$$

$$G_\lambda(\theta) = J_\lambda(\lambda) - \frac{TN\hat{u}_\lambda'(\theta)\Omega^{-1}\hat{u}(\theta)}{\hat{u}'(\theta)\Omega^{-1}\hat{u}(\theta)}, \tag{16}$$

where $\hat{u}_\lambda$ is the derivative of u with respect to λ evaluated at $\beta = \hat{\beta}(\theta)$. Note that the formulas for calculating $|\Omega|$ and Ω^{-1} introduced in Section 2.1 still apply.

For both Models defined in (1) and (10), standard errors of parameter estimates can be estimated using the sandwich estimator which we will discuss in detail in next section. The estimation procedures outlined above have been implemented using GAUSS CO or CML procedures with empirical data. It turns out that the above estimation procedures coupled with GAUSS CO work very well and convergence can be achieved quickly for both models. However, we find that when panel becomes large, computer memory problem may arise as there are many $TN \times TN$ matrices involved in the computation. In this case, it may be necessary to use a mainframe computer with a large memory to handle the computing work.

3 Asymptotic Properties of the QMLEs

In this section, we consider the asymptotic properties of the QMLEs. We consider the case where T is fixed and N goes large. Throughout, all quantities that are dependent on N are held implicitly. Quantities that are functions of some parameter(s), such as B, a function δ, and Ω, a function of ϕ and δ, are also held implicitly. Recall the parameter vector that needs to be estimated through an optimization process $\theta = \{\phi, \delta\}'$ for the standard model and $\theta = \{\phi, \delta, \lambda\}'$ for the transformed model. Let $\psi = \{\beta', \sigma_v^2, \theta'\}'$ be the full parameter vector. Let ψ_0 be the vector of true parameters and $\hat{\psi}$ its QMLE. A quantity evaluated at the true parameter values is denoted by adding a subscript '$_0$', e.g., B_0 is B evaluated at δ_0.

Let $\{h_N\}$ be a rate sequence that can be bounded or divergent as $N \to \infty$ such that the ratio $h_N/N \to 0$ as N goes to infinity. Some basic regularity conditions that are common to both models are listed below.

Assumption 1. *The $\{\mu_{0i}\}$ are iid with mean zero, variance $\sigma_{\mu 0}^2$, skewness $\alpha_{\mu 0}$ and centered kurtosis $\kappa_{\mu 0}$ (i.e., kurtosis minors 3); the $\{v_{0ti}\}$ are iid with mean zero, variance σ_{v0}^2, skewness α_{v0} and centered kurtosis κ_{v0}. The moments $E(|\mu_{0i}|^{4+\epsilon_1})$ and $E(|v_{0ti}|^{4+\epsilon_2})$ exist for some $\epsilon_1, \epsilon_2 > 0$.*

Assumption 2. *The elements w_{ij} of W are at most of order $O(h_N^{-1})$, uniformly in all i, j and $W_{ii} = 0$. The matrix B_0 is nonsingular. The sequence of matrices $\{W\}$ and $\{B_0^{-1}\}$ are uniformly bounded in both row and column sums.*

Assumption 3. *The $\{B^{-1}\}$ sequence are uniformly bounded in either row or column sums, uniformly in δ in a compact parameter space.*

Assumption 4. *The true θ_0 is the interior of a compact parameter space Θ.*

Assumption 5. *Define $\tilde{\ell}_{\max}(\theta) = \max_{\beta,\sigma_v^2} \mathrm{E}[\ell(\beta, \sigma_v^2, \theta)]$. The sequence $\{\tilde{\ell}_{\max}(\theta)\}$ has identifiably unique maximizers $\{\tilde{\theta}\}$, and $\tilde{\theta} \to \theta_0$ as N goes to infinity.*

Assumptions 1 spells out the essential features of the rescaled disturbances so that certain linear-quadratic forms in μ_0 or in v_0 obey the necessary probability

laws. Assumption 2 is originated by Lee (2004, Assumptions 2-5), which sets out the essential conditions for the weight matrix so that the systems (1) and (10) both have an equilibrium and Y or $h(Y,\lambda)$ has mean $X\beta$ or $X(\lambda)\beta$, and variance $\sigma_{v0}^2\Omega_0$. It also guarantees that this variance is bounded as N goes to infinity. Lee (2004) gives an extensive discussion on situations where this assumption is satisfied and on when h_N can be bounded and when it goes to infinity in a rate lower than N as N goes to infinity. Assumption 3 guarantees boundedness of certain matrices. Assumption 4 is standard. Assumption 5 is necessary for the parameters to be identifiable (White, 1994). Some additional assumptions are needed for each model.

3.1 Mixed effects model with spatial errors

An identifiability condition for the parameter vector β is necessary.

Assumption 6. *The elements of X are uniformly bounded constants for all N. The $\lim_{N\to\infty}\frac{1}{TN}X'\Omega^{-1}X$ exists and is nonsingular for all $\theta\in\Theta$.*

It is easy to see that $\mathrm{E}(\beta,\sigma_v^2.\theta) = -\frac{TN}{2}\log(\sigma_v^2) - \frac{1}{2}\log|\Omega| - \frac{1}{2\sigma_v^2}[\sigma_{v0}^2\mathrm{tr}(\Omega_0\Omega^{-1}) + (\beta_0-\beta)'X'\Omega^{-1}X(\beta_0-\beta)]$. Thus, the solution for the optimization problem defined in Assumption 5, $\tilde{\ell}_{\max}(\theta) = \max_{\beta,\sigma_v^2}\mathrm{E}[\ell(\beta,\sigma_v^2,\theta)]$, is

$$\begin{aligned}\tilde{\beta}(\theta) &= [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}X\beta_0 = \beta_0,\\ \tilde{\sigma}_v^2(\theta) &= \frac{1}{NT}\mathrm{E}\{[Y-X\tilde{\beta}(\theta)]'\Omega^{-1}[Y-X\tilde{\beta}(\theta)]\} = \frac{1}{NT}\sigma_{v0}^2\mathrm{tr}(\Omega_0\Omega^{-1}),\end{aligned}$$

which gives,

$$\tilde{\ell}_{\max}(\theta) = -\frac{TN}{2}\log[\tilde{\sigma}_v^2(\theta)] - \frac{1}{2}\log|\Omega| \tag{17}$$

Following White (1994, Theorem 3.4), consistency of $\hat{\theta}$ follows from the convergence of $\frac{1}{TN}[\ell_{\max}(\theta) - \tilde{\ell}_{\max}(\theta)] = \frac{1}{2}\{\log[\tilde{\sigma}_v^2(\theta)] - \log[\hat{\sigma}_v^2(\theta)]\}$ to zero in probability, uniformly on Θ. The consistency of $\hat{\beta}$ and σ_v^2 follows from that of $\hat{\theta}$.

Theorem 1. *Under Assumptions 1-6, $\hat{\psi}\xrightarrow{p}\psi_0$ as N goes to infinity.*

Asymptotic normality of the QMLE $\hat{\psi}$ can be derived from the Taylor expansion of $G(\hat{\psi}) = 0$ at ψ_0, which gives:

$$\sqrt{TN}(\hat{\psi} - \psi_0) = -\left(\frac{1}{TN}H(\bar{\psi})\right)^{-1}\frac{1}{\sqrt{TN}}G(\psi_0),$$

where $\bar{\psi} \xrightarrow{p} \psi_0$ as $\hat{\psi} \xrightarrow{p} \psi_0$, $G(\psi) = \partial\ell(\psi)/\partial\psi$, called the gradient function (which is score function when errors are normal), and $H(\psi) = (\partial^2/\partial\psi\partial\psi')\ell(\psi)$, called the Hessian matrix (which is the negative of the observed information matrix when errors are normal). The asymptotic normality of $\sqrt{TN}(\hat{\psi}-\psi_0)$ follows from the convergence of $\frac{1}{\sqrt{TN}}G(\psi_0)$ in law to normal, and the convergence of $\frac{1}{TN}H(\bar{\psi})^{-1}$ in probability. Define $\Phi_{10} = \Omega_0^{-1}(J_T \otimes I_N)\Omega_0^{-1}$ and $\Phi_{20} = \Omega_0^{-1}(I_T \otimes A_0)\Omega_0^{-1}$. From the log likelihood function $\ell(\psi)$ given in (3), we have the gradient function as

$$G(\psi_0) = \begin{cases} \frac{1}{\sigma_{v0}^2}X'\Omega_0^{-1}u_0, \\ \frac{1}{2\sigma_{v0}^4}u_0'\Omega_0^{-1}u_0 - \frac{TN}{2\sigma_{v0}^2}, \\ \frac{1}{2\sigma_{v0}^2}u_0'\Psi_{10}u_0 - \frac{1}{2}\mathrm{tr}(\Psi_{10}\Omega_0^{-1}), \\ \frac{1}{2\sigma_{v0}^2}u_0'\Psi_{20}u_0 - \frac{1}{2}\mathrm{tr}(\Psi_{20}\Omega_0^{-1}), \end{cases} \tag{18}$$

where $u_0 = (1_T \otimes I_N)\mu_0 + (I_T \otimes B_0^{-1})v_0$. The elements of $G(\psi_0)$ are seen to be either linear or quadratic functions of μ_0 or v_0, with iid elements. Hence, the asymptotic distributions of $\frac{1}{\sqrt{TN}}G(\psi_0)$ can be derived from the central limit theorems for linear-quadratic forms in Kelejian and Prochua (2001).

Theorem 2. *Under Assumptions 1-6, assume further that $\frac{1}{TN}(\partial/\partial\psi)H(\psi)$ is bounded in probability uniformly in a neighborhood of ψ_0. Then, we have*

$$\sqrt{TN}(\hat{\psi} - \psi_0) \xrightarrow{D} N(0, \Sigma_0^{-1}\Pi_0\Sigma_0^{-1})$$

as $N \to \infty$, where $\Pi_0 = \lim_{N\to\infty}\frac{1}{TN}\mathrm{Var}[G(\psi_0)]$ and $\Sigma_0 = -\lim_{N\to\infty}\frac{1}{TN}\mathrm{E}[H(\psi_0)]$, both assumed to exist, and Σ_0 is nonsingular.

Practical applications of Theorem 2 from inference point of view require both Π_0 and Σ_0 be estimated consistently and perhaps conveniently. Following lemma

provides a convenient tool for deriving the explicite expressions for $\mathrm{Var}[G(\psi_0)]$ and $\mathrm{E}[H(\psi_0)]$, based on which consistent estimates of Π_0 and Σ_0 can be obtained easily.

Lemma 1. *Let $\varepsilon = R_1 v_1 + R_2 v_2$ where v_1 and v_2 are two independent random vectors of iid elements such that v_{1i} has mean zero, variance σ_1^2, skewness α_1, and centered kurtosis κ_1, and that v_{2i} has mean zero, variance σ_2^2, skewness α_2, and centered kurtosis κ_2. R_1 and R_2 are two fixed matrices. Let C and D be two square matrices.*

(a) $\Omega = \frac{1}{\sigma_2^2}\mathrm{E}(\varepsilon\varepsilon') = \phi R_1 R_1' + R_2 R_2'$, *where* $\phi = \sigma_1^2/\sigma_2^2$,

(b) $E(\varepsilon' C \varepsilon) = \sigma_2^2 \mathrm{tr}(\Omega C)$,

(c) $g(C) = \frac{1}{\sigma_2^3}\mathrm{Cov}(\varepsilon, \varepsilon' C \varepsilon) = \phi^{\frac{3}{2}}\alpha_1 R_1 c_{11} + \alpha_2 R_2 c_{22}$,

(d) $f(C,D) = \frac{1}{\sigma^4}\mathrm{Cov}(\varepsilon' C\varepsilon, \varepsilon' D\varepsilon) = \phi^2\kappa_1 c_{11}' d_{11} + \kappa_2 c_{22}' d_{22} + 2\mathrm{tr}(\Omega C \Omega D)$,

where $c_{ij} = \mathrm{diag}(R_i' C R_j)$, and $d_{ij} = \mathrm{diag}(R_i' D R_j)$, $i, j = 1, 2$. When v_1 and v_2 are both normal, $g(\varepsilon, C) = 0$, and $f(\varepsilon, C, D) = 2\mathrm{tr}(\Omega C \Omega D)$.

The most useful result in Lemma 1 is that in part (d). Applying Lemma 1 on the elements of $G(\psi_0)$ with $\varepsilon = u_0$, $R_1 = 1_T \otimes I_N$, $R_2 = I_T \otimes B_0^{-1}$, $v_1 = \mu$, $v_2 = v$, $\sigma_1^2 = \sigma_{\mu 0}^2$, $\sigma_2^2 = \sigma_{v0}^2$, and $\phi = \sigma_{\mu 0}^2/\sigma_{v0}^2$, one obtains immediately,

$$\mathrm{Var}[G(\psi_0)] =$$

$$\left\{\begin{array}{llll} \frac{1}{\sigma_{v0}^2}X'\Omega_0^{-1}X, & \frac{1}{2\sigma_{v0}^3}X'\Omega_0^{-1}g(\Omega_0^{-1}), & \frac{1}{2\sigma_{v0}}X'\Omega_0^{-1}g(\Phi_{10}), & \frac{1}{2\sigma_{v0}}X'\Omega_0^{-1}g(\Phi_{20}) \\ \sim, & \frac{1}{4\sigma_{v0}^4}f(u_0, \Omega_0^{-1}, \Omega_0^{-1}) & \frac{1}{4\sigma_{v0}^2}f(\Omega_0^{-1}, \Phi_{10}) & \frac{1}{4\sigma_{v0}^2}f(\Omega_0^{-1}, \Phi_{20}) \\ \sim, & \sim, & \frac{1}{4}f(\Phi_{10}, \Phi_{10}) & \frac{1}{4}f(\Phi_{10}, \Phi_{20}) \\ \sim, & \sim, & \sim, & \frac{1}{4}f(\Phi_{20}, \Phi_{20}) \end{array}\right\} \quad (19)$$

With the exact expression of $H(\psi)$ (given in Appendix A), and Lemma 1(b), one easily obtain the expected negative Hessian matrix as

$$-\mathrm{E}[H(\psi_0)] = \left\{\begin{array}{llll} \frac{1}{\sigma_{v0}^2}X'\Omega_0^{-1}X, & 0_k, & 0_k, & 0_k \\ \sim, & \frac{TN}{2\sigma_{v0}^4}, & \frac{1}{2\sigma_{v0}^2}\mathrm{tr}(\Phi_{10}\Omega_0) & \frac{1}{2\sigma_{v0}^2}\mathrm{tr}(\Phi_{20}\Omega_0) \\ \sim, & \sim, & \frac{1}{2}\mathrm{tr}(\Phi_{10}\Omega_0\Phi_{10}\Omega_0) & \frac{1}{2}\mathrm{tr}(\Phi_{10}\Omega_0\Phi_{20}\Omega_0) \\ \sim, & \sim, & \sim, & \frac{1}{2}\mathrm{tr}(\Phi_{20}\Omega_0\Phi_{20}\Omega_0) \end{array}\right\} \quad (20)$$

Note that when errors are exact normal, $\alpha_{\mu 0} = \alpha_{v0} = \kappa_{\mu 0} = \kappa_{v0} = 0$. The fact that QMLE is robust against nonnormality is reflected by a non-zero values of $\alpha_{\mu 0}$, α_{v0}, $\kappa_{\mu 0}$ and κ_{v0}. In this case, the expression (19) reduces to (20). A practical issue now left is the way of estimating these quantities. Let $\hat{u} = Y - X\hat{\beta}$, and $\hat{B}$ be the QMLE of B_0. Run a GLS regression of $\hat{u}$ on R_1 with weights $\hat{R}_2 = I_T \otimes \hat{B}^{-1}$. Let $\hat{\mu}$ be the GLS estimate of the regression coefficients and $\hat{v}$ be the GLS estimates of the standardized residuals. Then, the skewness and centered kurtosis of $\hat{\mu}$ gives estimates of $\alpha_{\mu 0}$ and $\kappa_{\mu 0}$, and the skewness and centered kurtosis of $\hat{v}$ gives estimates of α_{v0} and κ_{v0}. Thus, the variance of the QMLE $\hat{\psi}$ can be conveniently estimated by

$$\widehat{\mathrm{Var}(\hat{\psi})} = \{\mathrm{E}[H(\psi_0)]\}^{-1}\mathrm{Var}[G(\psi_0)]\{\mathrm{E}[H(\psi_0)]\}^{-1}|_{\psi_0=\hat{\psi}}.$$

3.2 Transformed mixed effects model with spatial error

For the transformed mixed effects model with spatial error, the parameter vector θ that needs to be estimated through an optimization process (Equation (13)) contains an additional element, the transformation parameter λ. For ease of exposition, we use the same set of notation as in Section 3.1 but keep in mind that this extra element is involved everywhere, e.g., the gradient function $G(\psi)$ and the Hessian function $H(\psi)$. Also, the X matrix should be replaced everywhere by $X(\lambda)$.

Assumption 6*. *The elements of $X(\lambda)$ are uniformly bounded for all N, uniformly in λ in a compact set. The limit, $\lim_{N\to\infty} \frac{1}{TN} X'(\lambda)\Omega^{-1}X(\lambda)$, exists and is nonsingular for all $\theta \in \Theta$.*

It is easy to show that the optimal solution to the maximization problem defined in Assumption 5, $\tilde{\ell}_{\max}(\theta) = \max_{\beta,\sigma_v^2} \mathrm{E}[\ell(\beta, \sigma_v^2, \theta)]$, becomes

$$\begin{aligned}
\tilde{\beta}(\theta) &= [X'(\lambda)\Omega^{-1}X(\lambda)]^{-1}X'(\lambda)\Omega^{-1}\mathrm{E}[h(Y,\lambda)] \\
\tilde{\sigma}_v^2(\theta) &= (TN)^{-1}\mathrm{E}\{[h(Y,\lambda) - X(\lambda)\tilde{\beta}(\theta)]'\Omega^{-1}[h(Y,\lambda) - X(\lambda)\tilde{\beta}(\theta)]\},
\end{aligned}$$

which leads to $\tilde{\ell}_{\max}(\theta) = -\frac{TN}{2}\log[\tilde{\sigma}_v^2(\theta)] - \frac{1}{2}\log|\Omega| + \mathrm{E}[J(\lambda)]$, and

$$\frac{1}{TN}[\hat{\ell}_{\max}(\theta) - \tilde{\ell}_{\max}(\theta)] = -\frac{1}{2}[\log\hat{\sigma}_v^2(\theta) - \log\tilde{\sigma}_v^2(\theta)] + \frac{1}{TN}\{J(\lambda) - \mathrm{E}[J(\lambda)]\}.$$

Let $Y(\theta) = \Omega^{-\frac{1}{2}}h(Y,\lambda)$, and $P(\theta) = \Omega^{-\frac{1}{2}}X(\lambda)[X(\lambda)\Omega^{-1}X'(\lambda)]^{-1}X'(\lambda)\Omega^{-\frac{1}{2}}$. With the identification condition in Assumption 6a and the convergence of $\frac{1}{TN}[\hat{\ell}_{\max}(\theta) - \tilde{\ell}_{\max}(\theta)]$ in probability to zero uniformly on Θ, consistency of the QMLE $\hat{\psi}$ follows.

Theorem 3. *Under Assumptions 1-5 and Assumption 6*, assume further that (i)* $\|Y(\theta)\|^2 - \mathrm{E}\|Y(\theta)\|^2 = o_p(N)$ *where* $\|\cdot\|$ *is the Euclidean norm, (ii)* $\|P(\theta)Y(\theta)\|^2 = o_p(N)$*, (iii)* $\|P(\theta)\mathrm{E}[Y(\theta)]\|^2 = o(N)$*, and (iv)* $1'_{TN}[\log h_Y(Y,\lambda) - \log \mathrm{E}h_Y(Y,\lambda)] = o_p(N)$*, all uniformly on* Θ*. Then, we have,* $\hat{\psi} \xrightarrow{p} \psi_0$*, as* $N \to \infty$*.*

The treatment for the asymptotic normality of the QMLE of the transformed mixed effect model requires some additional assumptions and approximations. The gradient function is that of model (1) given in (18) after replacing X by $X(\lambda)$, plus the following additional element that corresponds to the transformation parameter λ,

$$G_\lambda(\psi_0) = J_\lambda(\lambda_0) - \frac{1}{\sigma_{v0}^2}u'_{0\lambda}\Omega^{-1}u_0, \tag{21}$$

where $J_\lambda(\lambda_0) = (d/d\lambda_0)J(\lambda_0)$ and $u_{0\lambda} = (\partial/\partial\lambda_0)u_0$. This can neither be written in linear forms nor in quadratic forms of u_0. Hence, the central limit theorems for linear-quadratic forms in Kelejian and Prucha (2001) cannot be directly applied. However, as we see below, under certain conditions it can be approximated by a linear-quadratic form. We consider the case where h is the Box and Cox (1964) power transformation:

$$h(y,\lambda) = \begin{cases} \frac{1}{\lambda}(y^\lambda - 1), & \lambda \neq 0, \\ \log y, & \lambda = 0, \end{cases} \quad y > 0. \tag{22}$$

Its first derivative has the form

$$h_\lambda(y,\lambda) = \begin{cases} \frac{1}{\lambda}[1 + \lambda h(y,\lambda)]\log y - \frac{1}{\lambda}h(y,\lambda), & \lambda \neq 0, \\ \frac{1}{2}(\log y)^2, & \lambda = 0. \end{cases}$$

In this case, we have $J(\lambda) = \sum_{t=1}^{T}\sum_{i=1}^{N} \log Y_{it}$. Define

$$\Delta_0 = \max_{t,i} \left| \frac{\lambda_0 \sqrt{\mathrm{Var}(u_{0,ti})}}{1 + \lambda_0 x'_{ti}(\lambda_0)\beta_0} \right|.$$

If $\Delta_0 \ll 1$, then some simple approximations to $\log Y$ as well as $h_\lambda(Y, \lambda_0)$ can be developed which enable us to write the gradient function (22) in a linear-quadratic form of u_0. This assumption falls in spirit into the framework of small-σ asymptotics of Bickel and Doksum (1981). Under this assumption, we have through a first-order Taylor series approximation,

$$\lambda_0 \log Y \approx \log(1_{TN} + \lambda_0\mu_0) + \lambda_0(1_{TN} + \lambda_0\mu_0)^{-1}u_0,$$

where $\mu_0 = X(\lambda_0)\beta_0$, and the log and inverse functions applied to $(1_{TN} + \lambda_0\mu_0)$ are operated elementwise. This leads to an approximation to the derivative of u_0,

$$u_{0\lambda} \approx a_0 + b_0 \odot u_0,$$

where $\odot$ denotes the Hadamard product, i.e., the elementwise multiplication, $a_0 = \frac{1}{\lambda_0^2}(1_{TN} + \lambda_0\mu_0) \odot \log(1_{TN} + \lambda_0\mu_0) - \frac{1}{\lambda_0}\mu_0 - X_\lambda(\lambda_0)\beta_0$, and $b_0 = \frac{1}{\lambda_0}\log(1_{TN} + \lambda_0\mu_0)$. Hence the gradient function corresponding to λ has the following approximation,

$$G_\lambda(\psi_0) \approx 1'_{TN}b_0 + \eta'_0 u_0 - \sigma_{v0}^{-2}u'_0\Phi_{30}u_0, \tag{23}$$

where $\Phi_{30} = \mathrm{diag}\{b_0\}\Omega_0^{-1}$ and $\eta_0 = (1_{TN} + \lambda_0\mu_0)^{-1} - \sigma_{v0}^{-2}\Omega_0^{-1}a_0$. Using Lemma 1, this leads immediately to approximations to the λ-related elements of $\mathrm{Var}[G(\psi_0)]$:

$$\mathrm{E}[G(\psi_0)G_\lambda(\psi_0)] \approx \begin{cases} X'(\lambda_0)\left(\eta_0 - \frac{1}{\sigma_{v0}}\Omega_0^{-1}g(\Psi_{30})\right), \\ \frac{1}{2\sigma_{v0}}\eta'_0 g(\Omega_0^{-1}) - \frac{1}{2\sigma_{v0}^2}f(\Omega_0^{-1}, \Phi_{30}), \\ \frac{\sigma_{v0}}{2}\eta'_0 g(\Phi_{10}) - \frac{1}{2}f(\Phi_{10}, \Phi_{30}), \\ \frac{\sigma_{v0}}{2}\eta'_0 g(\Phi_{20}) - \frac{1}{2}f(\Phi_{20}, \Phi_{30}), \\ \sigma_{v0}^2\eta'_0\Omega_0\eta_0 + f(\Phi_{30}, \Phi_{30}) - 2\sigma_{v0}\eta'_0 g(\Phi_{30}). \end{cases} \tag{24}$$

This together with (19) give the full expression of $\mathrm{Var}[G(\psi_0)]$. The full expressions for the gradient function $G(\psi)$ and the Hessian function $H(\psi)$ are given in the Appendix

A. Approximations to the expectations of the λ-related elements in the Hessian matrix are possible, though complicated, to give $\mathrm{E}[H(\psi_0)]$, but not necessary. As long as the Hessian matrix obeys some asymptotic properties as in the following theorem, one can use $-\frac{1}{TN}H(\hat{\psi})$ to estimate Σ_0 defined therein.

Theorem 4. *Under Assumptions 1-5 and Assumption 6*, assume further that (i) h is the Box-Cox power transformation and $\Delta_0 \ll 1$, (ii) $(\partial/\partial\psi)\frac{1}{TN}H(\psi)$ is bounded in probability, uniformly in a neighborhood of ψ, and (iii) $\frac{1}{TN}[H(\psi)-\mathrm{E}(H(\psi))] = o_p(1)$, uniformly in a neighborhood of ψ. We have,*

$$\sqrt{TN}(\hat{\psi}-\psi_0) \longrightarrow^D N(0, \Sigma_0^{-1}\Pi_0\Sigma_0^{-1}),$$

as $N \to \infty$, where $\Pi_0 = \lim_{N\to\infty}\frac{1}{TN}\mathrm{Var}[G(\psi_0)]$ and $\Sigma_0 = -\lim_{N\to\infty}\frac{1}{TN}\mathrm{E}[H(\psi_0)]$ both assumed to exist, and Σ_0 is nonsingular.

In practical applications, Σ_0 can be estimated consistently by $-\frac{1}{TN}H(\hat{\psi})$, and Π_0 can be estimated consistently by $\frac{1}{TN}\mathrm{Var}[G(\psi_0)]|_{\psi_0=\hat{\psi}}$, in which the skewness and kurtosis of μ and the skewness and kurtosis of v are estimated using the same method as for Model (1).

Unlike the case of the usual spatial panel models, estimation of the variance of gradient function seems to be one of the key issues in implementing the QMLE method for the transformed spatial panel model. The explicit expression of $\mathrm{Var}[G(\psi_0)]$ is not available and alternative methods or approximations have to be followed. Conventional methods include OPG (outer product of gradients) (Davidson and MacKinnon, 1993) which requires that the gradient can be written as a summation of TN independent elements, and the resampling method (Foster et al., 2001) which requires that the gradient function can be written as a U-statistics. However, neither is the case for our gradient function. Hence, an approximation method is followed.

4 An Empirical Illustration

In this section, we consider a numerical example to illustrate the model and the ML estimation procedure developed earlier. In particular, we consider the demand equations for cigarettes for United States, based on a panel of 46 states over 30 time periods (1963-1992), given as `CIGAR.TXT` on the Wiley web site associated with book of Baltagi (2001). The response variable Y = Cigarette sales in packs per capita. The covariates are X_1 = Price per pack of cigarettes; X_2 = Population; X_3 = Population above the age of 16; X_4 = Consumer price index with (1983=100); X_5 = Per capita disposable income; and X_6 = Minimum price in adjoining states per pack of cigarettes.

Earlier studies regarding states demand for cigarettes include Hamilton (1972), Baltagi and Levin (1986, 1992), Baltagi, Griffin and Xiong (2000), and Baltagi and Li (2004). Only Baltagi and Li (2004) has considered spatial effects in modeling the cigarettes demand, where some general explanations are given on why and how spatial correlation may arise in the demand for cigarettes.

Following Baltagi, et al. (2000), we treat the time periods effects as fixed. Corresponding to the major policy interventions in 1965, 1968 and 1971, we used a single dummy for each of the three multi-year periods: 1963-1964, 1965-1967 and 1968-1970, all inclusive, and a dummy for each of the rest of the years except the last year 1992, which is dropped out to prevent over parameterization. A similar treatment was given in Baltagi and Levin (1986). We consider fitting of three models: (I) both response and covariates are log transformed; (II) response is Box-Cox transformed, and covariates are log transformed; and (III) both response and covariates are Box-Cox transformed. Model (I) is similar to that of Baltagi and Li (2004) where they consider the prediction problem. It is reasonable to expect that our transformation model will give a better predictive performance when the transformation parameter is significantly different from zero or one, which are exactly the cases for the cigarette demand data. For each of the three models, two cases are considered: the case of without time effects and the case with time effects. For the spatial weighting matrix

W, we follow the first-order rook's contiguity relations. See Kelejian and Robinson (1995) for a good discussion on the spatial weighing matrix.

The estimation results are summarized in Table 1. Following general observations are in order: (i) spatial effect is strongly significant in all models considered, (ii) functional form is significantly different from the traditional log or linear forms, (iii) the individual random effects are also significant, and (iv) fixed time effects are highly significant collectively. We also found that the QMLE standard errors are larger than the MLE standard errors for most of the parameter estimates. As a result, the corresponding t-ratios are smaller.

It is interesting to note that three models give quite consistent estimates of spatial error correlation and cross-sectional random effects. Also, Model I & II (with or without fixed time effects) give very consistent estimates of transformation parameter λ. Model I is embedded in both Model II and Model III with λ specified as zero. The maximum values of the log likelihood function (without the constant) listed in the row labeled as `loglik` allows us to perform various likelihood ratio tests.[3] Likelihood ratio test (asymptotically χ^2 distributed with one degree of freedom) of Models I against II without fixed time effects results in a value of the test statistic 169.24, which becomes 160.80 when the pair of models with fixed time effects are compared. Similarly, the likelihood ratio test of Model I versus Model III has a test statistic value of 412.38 when time effects are absent and 504.82 when time effects are present. All tests strongly reject the null model. Thus, the conventional Cobb-Douglas functional form specification for the cigarette demand is strongly rejected by the data.

Comparison of the models (a) without fixed time effects versus (b) with fixed time effects shows the significance of the fixed time effects collectively. The likelihood ratio test of the Model I(a) versus Model I(b) has a statistic value of 89.76, of Model II(a)

[3]These tests need to be modified when errors are not exact normal. The distribution of the quasi-likelihood ratio test may be obtainable using the method described in Carroll et al. (1995). Alternatively, one may simply use the LM test with the QMLE variance-covariance estimates.

versus Model II(b), 81.42, and of Model III(a) versus Model III(b), 182.2. According to the values of the maximized log likelihood function, Model III with fixed time effects fits the data the best. Thus, it can be used to perform predictions and to carry out various other tasks. It is reasonable to believe that this model performs better in its predictive performance than that considered in Baltagi and Li (2004).

Price (X_4) and income (X_5) elasticities are often of interest. Model I specifies that these are constant. For price elasticity, it is estimated to be -1.1699 by Model I(a) and -1.1707 by Model I(b); for income elasticity, it is estimated to be 0.1147 by Model I(a) and 0.4606 by Model I(b). However, Model I is incorrectly specified, hence this constant elasticity assumption is subject to question. The elasticity of a covariate x relative to a response y in a general transformation model takes the form:

$$E_{y|x} = \beta \frac{X}{Y} \cdot \frac{g_x(x)}{h_y(y)}$$

where $g(x)$ and $h(y)$ are, respectively, transformations applied to x and y, and $g_x(x)$ and $h_y(y)$ are the derivatives. So, in the case of applying Box-Cox transformations on both sides, we have $E_{y|x} = \beta(x/y)^\lambda$, and in the case of applying Box-Cox transformation on y and log transformation on x, we have $E_{y|x} = \beta/y^\lambda$.

Hence, in both Models II and III, the elasticity is not constant, which depends on the y value in Model II and on both x and y values in Model III. From Model III, the price elasticity at the mean levels (average sale 123.95 and average price 68.70) is estimated to be $\hat{E}_{y|x} = \hat{\beta}_4(\bar{X}_4/\bar{Y})^{\hat{\lambda}} = -1.7156(68.70/123.95)^{-0.5262} = -2.3403$ using Model III(a) and -2.4488 using Model III(b). The same numbers for income elasticity at the means (123.95 and 7525) are -0.0079 and -0.0511. Similarly, the price elasticity estimates at the mean sales level estimated from Model II, without or with fixed time effects, are $\hat{E}_{y|x} = \hat{\beta}_4/\bar{Y}^{\hat{\lambda}} = (-0.4203, -0.3222)$; and the corresponding income elasticity estimates are $(0.1172, -0.0143)$.

Per capita sales vary a lot from state to state. According to Models II and III, the effects of covariates on sales can be quite different at different levels of cigarette sales. For example, for states with per capita sales 85 packs, the price elasticity

at the average price level becomes $(-0.3262, -0.2513)$ and $(-1.9190, -2.0013)$, respectively, from Models II(a, b) and III(a, b); and for states with per capita sales 300 packs, the price elasticity at average price level becomes $(-0.7610, -0.5765)$ and $(-3.7262, -3.9290)$ from Models II(a, b) and III(a, b), respectively. This shows that price has much bigger impact on sales when sale volume is high than when it is low. Similar conclusions apply to income variable, as well as other covariates.

5 Conclusions

We have introduced two important techniques, quasi-maximum likelihood estimation and data transformation, into the modelling of spatial panel data. For spatial regression models, many authors have advocated the use of maximum likelihood method for model estimation (e.g., Anselin, 1988; Elhorst, 2003). Quasi-maximum likelihood estimation provides robust standard error estimates, which makes the likelihood-based method more attractive. Data transformation aims to bring the data to near normality, which makes normality-based QMLE more valid. Empirical results show the importance of applying both techniques in modelling the spatial panel data. Some immediate future work may be assessing the predictive performance of the models and more empirical applications.

Appendix A: The Gradient and Hessian

We first present the gradient and Hessian functions for Model (1). The same functions for the transformed model, Model (10), can be obtained by adding the λ-related elements on the gradient and Hessian matrix for Model (1). The elements of the gradient vector $G(\psi)$ for Model (1) are,

$$\begin{aligned}
G_\beta &= \frac{\partial \ell}{\partial \beta} = \tfrac{1}{\sigma_v^2} X'\Omega^{-1}u \\
G_{\sigma_v^2} &= \frac{\partial \ell}{\partial \sigma_v^2} = \tfrac{1}{2\sigma_v^4} u'\Omega^{-1}u - \tfrac{NT}{2\sigma_v^2} \\
G_\phi &= \frac{\partial \ell}{\partial \phi} = \tfrac{1}{2\sigma_v^2} u'\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}u - \tfrac{1}{2}\mathrm{tr}[\Omega^{-1}(J_T \otimes I_N)] \\
G_\delta &= \frac{\partial \ell}{\partial \delta} = \tfrac{1}{2\sigma_v^2} u'\Omega^{-1}(I_T \otimes A)\Omega^{-1}u - \tfrac{1}{2}\mathrm{tr}[\Omega^{-1}(I_T \otimes A)]
\end{aligned}$$

and the elements of the Hessian matrix $H(\psi)$ for Model (1),

$$\begin{aligned}
H_{\beta\beta} &= -\tfrac{1}{\sigma_v^2} X'\Omega^{-1}X \\
H_{\beta\sigma_v^2} &= -\tfrac{1}{\sigma_v^4} X'\Omega^{-1}u \\
H_{\beta\phi} &= -\tfrac{1}{\sigma_v^2} X'\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}u \\
H_{\beta\delta} &= -\tfrac{1}{\sigma_v^2} X'\Omega^{-1}(I_T \otimes A)\Omega^{-1}u \\
H_{\sigma_v^2\sigma_v^2} &= \tfrac{NT}{2\sigma_v^4} - \tfrac{1}{\sigma_v^6} u'\Omega^{-1}u \\
H_{\sigma_v^2\phi} &= -\tfrac{1}{2\sigma_v^4} u'\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}u \\
H_{\sigma_v^2\delta} &= -\tfrac{1}{2\sigma_v^4} u'\Omega^{-1}(I_T \otimes A)\Omega^{-1}u \\
H_{\phi\phi} &= \tfrac{1}{2}\mathrm{tr}[\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}(J_T \otimes I_N)] - \tfrac{1}{\sigma_v^2} u'\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}u \\
H_{\phi\delta} &= \tfrac{1}{2}\mathrm{tr}[\Omega^{-1}(I_T \otimes A)\Omega^{-1}(J_T \otimes I_N)] - \tfrac{1}{\sigma_v^2} u'\Omega^{-1}(I_T \otimes A)\Omega^{-1}(J_T \otimes I_N)\Omega^{-1}u \\
H_{\delta\delta} &= \tfrac{1}{2}\mathrm{tr}[\Omega^{-1}(I_T \otimes A)\Omega^{-1}(I_T \otimes A) - \Omega^{-1}(I_T \otimes \tfrac{\partial A}{\partial \delta})] \\
&\quad -\tfrac{1}{\sigma_v^2} u'\Omega^{-1}(I_T \otimes A)\Omega^{-1}(I_T \otimes A)\Omega^{-1}u + \tfrac{1}{2\sigma_v^2} u'\Omega^{-1}(I_T \otimes \tfrac{\partial A}{\partial \delta})\Omega^{-1}u
\end{aligned}$$

where A is given in (6) and $\frac{\partial A}{\partial \delta} = 2(B'B)^{-1}[(W'B + B'W)A - W'W]$. For the gradient function of Model (10), change X to $X(\lambda)$ and add the following element

$$G_\lambda = \frac{\partial \ell}{\partial \lambda} = J_\lambda(\lambda) - \frac{1}{\sigma_v^2} u_\lambda' \Omega^{-1} u,$$

and for the Hessian of Model (10), replace X by $X(\lambda)$ and add the following elements

$$\begin{aligned}
H_{\beta\lambda} &= \tfrac{1}{\sigma_v^2}[X'_\lambda(\lambda)\Omega^{-1}u + X'(\lambda)\Omega^{-1}u_\lambda] \\
H_{\sigma_v^2\lambda} &= \tfrac{1}{\sigma_v^4}u'_\lambda\Omega^{-1}u \\
H_{\phi\lambda} &= \tfrac{1}{\sigma_v^2}u'_\lambda\Omega^{-1}(J_T\otimes I_N)\Omega^{-1}u \\
H_{\delta\lambda} &= \tfrac{1}{\sigma_v^2}u'_\lambda\Omega^{-1}(I_T\otimes A)\Omega^{-1}u \\
H_{\lambda\lambda} &= -\tfrac{1}{\sigma_v^2}(u'_{\lambda\lambda}\Omega^{-1}u + u'_\lambda\Omega^{-1}u_\lambda) + J_{\lambda\lambda}(\lambda)
\end{aligned}$$

Appendix B: Proofs of the Theorems

Proof of Theorem 1: Consistency of $\hat{\psi}$ follows from the uniform convergence of $\frac{1}{TN}[\ell_{\max}(\theta) - \tilde{\ell}_{\max}(\theta)]$ to zero on Θ and the uniqueness identification condition given in Assumption 5 (White, 1994, Theorem 3.4.). Since $\frac{1}{TN}[\ell_{\max}(\theta) - \tilde{\ell}_{\max}(\theta)] = -\frac{1}{2}[\log\hat{\sigma}_v^2(\theta) - \log\tilde{\sigma}_v^2(\theta)]$. It suffices to show that

$$\hat{\sigma}_v^2(\theta) - \tilde{\sigma}_v^2(\theta) \xrightarrow{p} 0, \text{ uniformly on } \Theta.$$

Now, $\tilde{\sigma}_v^2(\theta) = \frac{1}{TN}\text{tr}(\Omega_0^{-\frac{1}{2}}\Omega^{-1}\Omega_0^{-\frac{1}{2}})$, and

$$\hat{\sigma}_v^2(\theta) = \frac{1}{TN}[Y - X\hat{\beta}(\theta)]'\Omega^{-1}[Y - X\hat{\beta}(\theta)] = \frac{1}{TN}Y'\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}Y,$$

where $M = I_{TN} - \Omega^{-\frac{1}{2}}X(X'\Omega^{-1}X)^{-1}X'\Omega^{-\frac{1}{2}}$, a projection matrix projecting onto a space orthogonal to the space spanned by the columns of $\Omega^{-\frac{1}{2}}X$. Define $u_0^* = \Omega_0^{\frac{1}{2}}u_0$. We have,

$$\hat{\sigma}_v^2(\theta) = \frac{1}{TN}u_0'\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}u_0 = \frac{1}{TN}u_0^{*\prime}\Omega_0^{\frac{1}{2}}\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}\Omega_0^{\frac{1}{2}}u_0^*,$$

which gives

$$\begin{aligned}
\hat{\sigma}_v^2(\theta) - \tilde{\sigma}_v^2(\theta) &= \frac{1}{TN}u_0^{*\prime}\Omega_0^{\frac{1}{2}}\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}\Omega_0^{\frac{1}{2}}u_0^* - \frac{1}{TN}\text{tr}(\Omega_0^{\frac{1}{2}}\Omega^{-1}\Omega_0^{\frac{1}{2}}) \\
&= \frac{1}{TN}\left[u_0^{*\prime}\Omega_0^{\frac{1}{2}}\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}\Omega_0^{\frac{1}{2}}u_0^* - \text{tr}(\Omega_0^{\frac{1}{2}}\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}\Omega_0^{\frac{1}{2}})\right] \\
&\quad + \frac{1}{TN}\left[\text{tr}(\Omega_0^{\frac{1}{2}}\Omega^{-\frac{1}{2}}M\Omega^{-\frac{1}{2}}\Omega_0^{\frac{1}{2}}) - \text{tr}(\Omega_0^{\frac{1}{2}}\Omega^{-1}\Omega_0^{\frac{1}{2}})\right].
\end{aligned}$$

It can be shown that the first term above is $o_p(1)$ and second term is $o(1)$, uniformly on Θ. Hence $\frac{1}{TN}[\ell_{\max}(\theta) - \tilde{\ell}_{\max}(\theta)] \xrightarrow{p} 0$ uniformly on Θ. This implies that $\hat{\theta} \xrightarrow{p} \theta_0$, and hence $\hat{\psi} \xrightarrow{p} \psi_0$.

Proof of Theorem 2: Proof of the theorem starts from the following Taylor expansion

$$0 = \frac{1}{\sqrt{TN}} G(\hat{\psi}) = \frac{1}{\sqrt{TN}} G(\psi_0) + \left(\frac{1}{TN} H(\bar{\psi})\right) \sqrt{TN}(\hat{\psi} - \psi_0).$$

As each element of the gradient vector can be written as linear, or quadratic, or linear-quadratic forms of μ_0 or v_0, Lindeberg-Feller central limit theorem can be used to prove the asymptotic normality of the elements that are of linear form, and Kelejian and Prucha (2001) central limit theorem can be used to prove the asymptotic normality of the components that are quadratic or linear-quadratic functions of μ_0 or v_0. Finally, Cramer-Wold device can be used to prove the joint asymptotic normality of the gradient vector, as linear combinations of linear and quadratic functions of μ_0 are a linear-quadratic function of μ_0. The same applies to v_0. Thus,

$$\frac{1}{TN} G(\psi_0) \longrightarrow^D N(0, \Pi_0)$$

where $\Pi_0 = \lim_{N\to\infty} \frac{1}{TN} \mathrm{E}[G'(\psi_0) G(\psi_0)]$. By the mean value theorem, we have

$$\frac{1}{TN} H(\bar{\psi}) - \frac{1}{TN} H(\psi_0) = \frac{1}{TN} (\partial/\partial\psi) H(\tilde{\psi})(\bar{\psi} - \psi_0) = o_p(1).$$

Finally, all the elements of the Hessian matrix (given in Appendix A) can be written as either linear, or quadratic functions of μ_0 or v_0. By showing these linear functions and quadratic functions divided by TN are all $O_p(1)$, we have

$$\frac{1}{TN} H(\psi_0) - \frac{1}{TN} \mathrm{E}[H(\psi_0)] = o_p(1).$$

The result of the theorem follows.

Proof of Theorem 3: The $\hat{\sigma}_v^2(\theta)$ and $\tilde{\sigma}_v^2(\theta)$ can be written as

$$\hat{\sigma}_v^2(\theta) = \frac{1}{TN} Y(\theta)'[I_{TN} - P(\theta)] Y(\theta) = \frac{1}{TN} Y(\theta)'Y(\theta) - \frac{1}{TN} Y(\theta)'P(\theta)Y(\theta), \text{ and}$$

$$\hat{\sigma}_v^2(\theta) = \frac{1}{TN}\mathrm{E}[Y(\theta)'Y(\theta)] - \frac{1}{TN}\mathrm{E}[Y(\theta)]'P(\theta)\mathrm{E}[Y(\theta)].$$

Hence,

$$\begin{aligned}\hat{\sigma}_v^2(\theta) - \tilde{\sigma}_v^2(\theta) &= \frac{1}{TN}\left\{Y(\theta)'Y(\theta) - \mathrm{E}[Y(\theta)'Y(\theta)]\right\} \\ &\quad -\frac{1}{TN}Y(\theta)'P(\theta)Y(\theta) + \frac{1}{TN}\mathrm{E}[Y(\theta)]'P(\theta)\mathrm{E}[Y(\theta)] \to^p 0.\end{aligned}$$

Further, $J(\lambda) - EJ(\lambda) = 1'_{TN}[\log h_Y(Y,\lambda) - \log Eh_Y(Y,\lambda)] = o(N)$ uniformly on Θ. We have $\frac{1}{TN}[\ell(\theta) - \tilde{\ell}(\theta)] \xrightarrow{p} 0$, uniformly on Θ. Consistency of $\hat{\theta}$ follows, which gives the consistency of $\hat{\beta}(\hat{\theta})$ and $\hat{\sigma}_v^2(\hat{\psi})$, and the consistency of the QMLE $\hat{\psi}$.

Proof of Theorem 4: As the gradient function specified in (18) and (23) can be either written as or approximated by linear or quadratic forms in μ_0 or in v_0. The proof of the theorem is similar to that of Theorem 2.

References

[1] Anselin, L. (2001). Spatial Econometrics. In B. H. Baltagi (Eds.) *A Companion to Theoretical Econometrics*. Blackwell.

[2] Anselin, L. and Bera, A. K. (1998). Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics. In A. Ulah and D. E. A. Giles (Eds.) *Handbook of Applied Economic Statistics*. New York: Marcel Dekker.

[3] Azomahou, T. (1999). Estimation of spatial panel data models using a minimum distance estimator: application. Working Papers of BETA, No. 9912.

[4] Azomahou, T. (2000). Semivariogram estimation and panel data structure in spatial models: an empirical analysis. *Paper No. 137, Computing in Economics and Finance 2000, Society for Computational Economics*.

[5] Azomahou, T. (2001). GMM Estimation of lattice models using panel data: application. Working Papers of BETA, No. 2001-09.

[6] Baltagi, B. H. (2001). *Econometric Analysis of Panel Data*. New York: John Wiley & Sons.

[7] Baltagi, B. H. and Levin, D. (1986). Estimating dynamic demand for cigarettes using panel data: The effects of bootlegging, taxation, and advertising reconsidered. *Review of Economics and Statistics* **68**, 148-155.

[8] Baltagi, B. H. and Levin, D. (1992). Cigarette taxation: Raising revenue and reducing consumption. *Structure Change and Economic Dynamics* **3**, 321-335.

[9] Baltagi, B. H. and Li D. (2001). LM tests for functional form and spatial correlation. *International Regional Science Review* **24**, 194-225.

[10] Baltagi, B. H. and Li, D. (2004). Prediction in the panel data model with spatial correlation. In L. Anselin, R. Florax and S. J. Rey (Eds.) *Advances in Spatial Econometrics*. Springer-Verlag: Berlin.

[11] Baltagi, B. H., Griffin, J. M. and Xiong, W. (2000). To pool or not to pool: Homogeneous versus heterogeneous estimators applied to cigarette demand. *Review of Economics and Statistics* **82**, 117-126.

[12] Baltagi, B. H., Song, S. H., and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics* **117**, 123-150.

[13] Baltagi, B. H., Song, S. H., and Jung, B. C. (2004). Testing for serial correlation, spatial autocorrelation and random effects using panel model. Working paper (also presented at ESAM2004 Melbourne).

[14] Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* **76**, 296-311.

[15] Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations (with discussion). *J. R. Statist. Soc.* **B 26**, 211-46.

[16] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement errors in nonlinear models.* Chapman & Hall: London.

[17] Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics.* Oxford University Press: Oxford.

[18] Elhorst, J. P. (2003). Specification and Estimation of Spatial Panel Data Models. *International Regional Science Review* **26**, 244-268.

[19] Foster, A. M., Tian, L. and Wei, L. J. (2001). Estimation for the Box-Cox Transformation Model without Assuming Parameter Error Distribution. *Journal of the American Statistical Association* **96**, 1097-1101.

[20] Griffith, D. A. (1988). *Advanced Spatial Statistics.* Dordrecht, the Netherlands: Kluwer.

[21] Griffith, D. A., Paelinck, J. H. P., and van Gastel, R. A. (1998). The Box-Cox transformation: computational and interpretation features of the parameters, in D. A. Griffith and C. Amrhein (Eds.) *Econometric Advances in Spatial Modelling and Methodology.* Amsterdam: Kluwer.

[22] Hamilton, J. L. (1972). The demand for cigarettes: advertising, the health care, and the cigarette advertising ban. *Review of Economics and Statistics* **54**, 401-411.

[23] Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distributions of the Moran I test statistic with applications. *Journal of Econometrics* **104**, 219-257.

[24] Kelejian, H. H. and Robinson, D. P. (1995). Spatial correlations: a suggested alternative to the autoregressive model. In L. Anselin and R. Florax (Eds.) *New Directions in Spatial Econometrics*. Berlin: Springer.

[25] Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**, 1899-1925.

[26] Magnus, J. R. (1982). Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. *Journal of Econometrics* **19**, 239-285.

[27] Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition*. John Wiley & Sons.

[28] Pace, R. K., Barry, R., Slawson Jr., V. C. and Sirmans, C. F. (2004). Simultaneous Spatial and Functional Form Transformation.In L. Anselin, R. Florax and S. J. Rey (Eds.) *Advances in Spatial Econometrics*. Springer-Verlag: Berlin.

[29] Pesearan, M. H. (2002). Estimation and inference in large heterogeneous panels with cross section dependence. *University of Cambridge DAE Working Paper No. 0305 and CESifo Working Paper Series No. 869.*

[30] Pesearan, M. H. (2004). General diagnostic tests for cross section dependence in panels. *University of Cambridge CERifo Working Paper No. 1229 and IZA Discussion Paper No. 1240.*

[31] van Gastel, R. A. and Paelinck, J. H. P. (1995). Computation of Box-Cox transform parameter: a new method and its applications to spatial econometrics, in L. Anselin and R. Florax (Eds.), *New Directions in Spatial Econometrics*, Springer-Verlag: Berlin.

[32] White, H. (1994). *Estimation, Inference and Specification Analysis.* New York: Cambridge University Press.

[33] White, H. (2001). *Asymptotic Theory for Econometricians.* New York: Academic Press.

Table 1. Estimation Results for the Cigarette Demand Data

Model	Par.	(a) Without Fixed Time Effects			(b) With Fixed Time Effects		
		Par. Est.	t-stat MLE	t-stat QMLE	Par. Est.	t-stat MLE	t-stat QMLE
I	β_0	2.4748	10.1095	10.3897	3.2262	3.9119	3.9208
	β_1	-0.9020	-26.9864	-26.9902	-1.0112	-25.2779	-25.3071
	β_2	0.5309	3.7139	3.7527	0.5260	3.4517	3.4942
	β_3	-0.5081	-3.6088	-3.6285	-0.5084	-3.3738	-3.4032
	β_4	0.0629	1.2364	1.2369	0.2000	1.0554	1.0572
	β_5	0.5448	13.3884	13.4010	0.5755	11.9703	11.9816
	β_6	0.1597	4.3794	4.3832	-0.0587	-1.0875	-1.0909
	σ_v	0.0731	349.4144	183.2402	0.0714	359.3930	192.0414
	ϕ	5.0560	4.4211	4.6783	5.1515	4.4670	4.6823
	δ	0.3535	14.4587	11.3845	0.2433	8.5873	7.3215
	loglik		-4756.67			-4711.79	
II	β_0	1.3431	18.8030	9.6844	1.3991	15.5916	8.7273
	β_1	-0.0345	-4.0950	-2.0934	-0.0401	-4.0167	-2.1004
	β_2	0.0085	1.3666	1.0072	0.0069	1.0239	0.8005
	β_3	-0.0072	-1.1943	-0.9044	-0.0059	-0.8913	-0.7113
	β_4	0.0020	0.9206	0.8545	-0.0003	-0.0316	-0.0313
	β_5	0.0214	3.9600	2.0805	0.0261	3.9974	2.1569
	β_6	0.0046	2.4890	1.6402	-0.0021	-0.9207	-0.8608
	σ_v	0.0027	766.9198	581.0148	0.0028	723.9371	405.9130
	ϕ	5.8541	4.3764	3.8569	5.8179	4.4389	4.0062
	δ	0.4530	18.9454	20.4597	0.3441	11.9208	11.9092
	λ	-0.6717	-13.4517	-6.8397	-0.6582	-13.0511	-6.7793
	loglik		-4672.05			-4631.34	
III	β_0	-7.6873	-12.5792	-7.2281	-8.2668	-13.1053	-8.1448
	β_1	-0.4476	-18.3674	-11.0216	-0.3797	-14.1564	-9.3236
	β_2	2.5704	7.9233	5.9917	2.5984	8.0228	6.1361
	β_3	-1.7156	-7.5380	-6.4724	-1.7859	-7.9064	-6.8378
	β_4	-0.0687	-1.9503	-1.7875	-0.4592	-4.7257	-4.6864
	β_5	4.6517	11.2584	6.1429	5.2974	12.1572	7.2700
	β_6	-0.0333	-1.7229	-1.6236	0.0482	2.0414	2.0291
	σ_v	0.0048	806.4606	187.2567	0.0044	847.1905	310.9054
	ϕ	13.8558	4.2189	3.4417	13.9944	4.2733	3.5204
	δ	0.5895	29.1003	26.1723	0.4001	13.5572	11.3791
	λ	-0.5262	-19.9278	-10.4136	-0.5349	-19.6440	-11.3226
	loglik		-4550.48			-4459.38	