Latent instrumental variables

Ebbes, P.

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 7

A Nonparametric Bayesian LIV approach

7.1 Introduction

In this chapter we introduce two important extensions of the standard LIV model presented in chapters 3 and 4. Firstly, we generalize the multinomial distribution of the latent instrument to a general distribution $G$. Secondly, we consider endogeneity in two commonly used multilevel models and we suggest how the method and results from this chapter may improve on the Hausman-Taylor approach discussed in the previous chapter. Besides, the method proposed here can be applied to a situation with more than one endogenous variable. We present a Bayesian framework that can be applied to a wide variety of models and allows for exact finite-sample inference. The methodological results and the simulation studies that we present are promising, yet further research is required and we suggest several steps for that.

The latent instrumental variable (LIV) model introduced in the previous chapters can be used to estimate linear regression models where one regressor is correlated with the error term. The LIV approach assumes the existence of a discrete instrument with unobserved category membership. Hence, the latent instrument has a multinomial distribution and the unconditional probability distribution of the observations $(y, x)$ is a mixture of $m$ bivariate normal distri-
bution functions, assuming normally distributed error-terms. Hence, the LIV method does not require the availability of observed instrumental variables to estimate the regression parameters if regressor-error correlations are suspected. This is an advantage over classical instrumental variables estimation since in many applications instruments are not available. Besides, available instruments may be of bad quality (weak or endogenous) and, hence, substantive conclusions for the same phenomenon may be different when other instruments are used (see e.g. chapters 2 and 5).

In this chapter we relax the assumption that the unobserved instrument has a multinomial distribution with \( m \) categories. Instead, we let the data determine the ‘best’ distribution \( G \) of the unobserved instrument. As such, this approach is potentially more efficient than the standard LIV model because the distribution of the instrument is fully estimated from the data and is not limited to an assumed multinomial distribution with \( m \) groups. Besides, the number of categories of the unobserved instrument is an unknown parameter that is determined by the data. So, no tests for the number of groups are required. Furthermore, we consider endogeneity issues in more general multilevel models. Ebbes, Böckenholt and Wedel (2004) put forward that endogeneity in multilevel models is more complex because of the presence of random components at various level. Traditional methods (fixed-effects estimation, random-effects estimation, Mundlak’s approach, or the Hausman-Taylor estimator) are shown to be limited in various ways and they present a list of open problems, some of which will be addressed here (see also chapter 6).

We propose a nonparametric Bayesian approach to estimate the regression parameters and the distribution of the unobserved instrument simultaneously. Nonparametric Bayes models have originally been proposed to alleviate the parametric assumptions often made in standard hierarchical linear models (Escobar, 1994, Escobar and West, 1998, Ibrahim and Kleinman, 1998). At the heart of hierarchical models are assumptions on the distributions of various model parameters, which can often be questioned, and results are found to be sensitive to assumed forms of the distributions. Nonparametric Bayesian mod-
7.1 Introduction

els provide a way to alleviate these parametric assumptions by using Dirichlet processes. A Dirichlet process is used as a prior distribution on the family of distributions, i.e. a prior distribution is specified on the space of all possible distribution functions. Before Escobar’s (1994) results on nonparametric Bayes models, applications of these methods were limited because of computational difficulties. However, he solved this by developing a MCMC sample algorithm that is fairly easy to implement. Since then several studies used nonparametric Bayes techniques in, for instance, density estimation or hierarchical modeling. Dey, Müller and Sinha (1998) give an overview of recent developments in this area.

One advantage of using a Bayesian nonparametric approach to model the distribution of the unobserved instrument is that it bypasses the need to determine the correct number of mixing components post hoc, while retaining the ability to recover a variety of distributions in a unified modeling framework (cf. Kim, Menzefricke and Feinberg, 2004). I.e., whatever the true distribution of the instruments is, the Bayes estimate converges to it (Ferguson, 1973). Antoniak (1974) shows that clustering is inherent to the Dirichlet process and there is a positive probability that the number of support points found is smaller than the number of sample points.

In section 7.2 we briefly introduce the nonparametric Bayes approach for a simple multilevel model with a latent instrument to solve for level-1 dependencies. We discuss model specification, the Dirichlet process prior, and we present an estimation scheme. Next, we extend this idea to a general hierarchical model where individual level covariates may be endogenous (i.e. level-2 endogeneity). We show that an unobserved instrument with a Dirichlet process can be incorporated in a similar way as for the simple multilevel model. We present simulation results in section 7.4 and the conclusions and a discussion of the results found are presented in section 7.5. Besides, we propose steps for further research.
7.2 A simple multilevel model with a general latent instrument

We consider the following simple multilevel model

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \]  

(7.1)

where \( i = 1, ..., n \), \( j = 1, ..., m_i \), and \( n_t = \sum_i m_i \) is the total number of observations. The error term \( \epsilon_{ij} \) has mean zero and variance \( \sigma^2 \). We do not assume that \( \text{E} \left( x_{ij} \epsilon_{ij} \right) \) equals zero. For the sake of simplicity, we omit further regressors and random terms, but these can be easily included since the Dirichlet process for the unobserved instrument is unaffected and MCMC estimation is conditional on all other parameters and observations. In the terminology of chapter 6, we consider here level-1 dependencies. Contrary to level-2 dependencies, level-1 dependencies are more difficult to address, since potential remedies require the availability of external instrumental variables that may not be available or may be of bad quality (subsection 6.5.1).

The endogeneity of \( x_{ij} \) is modeled as follows

\[
\begin{align*}
  y_{ij} &= \beta_0 + \beta_1 x_{ij} + \epsilon_{ij} \\
  x_{ij} &= \theta_{ij} + \nu_{ij},
\end{align*}
\]

(7.2)

where \( \theta_{ij} \) is the unobserved instrument and a ‘nuisance’ parameter. \( x_{ij} \) is endogenous when \( \text{E} \left( x_{ij} \epsilon_{ij} \right) \neq 0 \). We assume that the endogenous regressor \( x_{ij} \) can be split up in an exogenous part \( \theta_{ij} \) and an endogenous part \( \nu_{ij} \) where the latter is correlated with \( \epsilon_{ij} \). Furthermore, we assume that \( \epsilon_{ij} \) and \( \nu_{ij} \) have mean zero and variance-covariance matrix

\[
\Sigma = \begin{bmatrix}
\sigma^2 & \sigma_{\epsilon\nu} \\
\sigma_{\epsilon\nu} & \sigma^2
\end{bmatrix}.
\]

(7.3)

Instead of assuming a discrete distribution with \( k \) categories for \( \theta_{ij} \), we make a very general assumption and assume a Dirichlet process for the distribution
7.2 A simple multilevel model with a general latent instrument

of \(\theta_{ij}\) (Ferguson, 1973, Antoniak, 1974). To be more specific, the unobserved instruments are independently and identically distributed as \(G\), where we do not assume a specific parametric form for \(G\). Instead, it has a Dirichlet process prior, denoted by \(DP(\alpha, G_0)\), where \(\alpha > 0\) is a concentration parameter and \(G_0\) is the baseline prior distribution with density \(g_0\). A Dirichlet prior places a distribution (probability measure) on the space of all distribution functions for \(G\). For large values of \(\alpha\), \(G\) is very likely to be close to \(G_0\), whereas for small values of \(\alpha\), the mass of \(G\) is likely to be concentrated on a few atoms. In fact, the support of the Dirichlet process is the class of all distribution functions, and is, hence, very large. The nonparametric model allows the data to adapt a \(G\) that is skewed, has ‘shoulders’, is multimodal, or any general shape different from \(G_0\) (cf. MacEachern, 1998). We present the definition and some technical details on the Dirichlet process in appendix 7A. For more details on recent results, estimation, and applications, see Dey, Müller and Sinha (1998).

The distribution of \(\theta_{ij}\) is specified as follows

\[
\begin{align*}
\theta_{ij} | G & \sim G \quad (7.4) \\
G | \alpha, \lambda & \sim DP(\alpha, G_0(.|\lambda)),
\end{align*}
\]

where \(\theta_{ij}\)’s are independent and \(G_0\) is the baseline prior defined by the parameter \(\lambda\). As stated above, whatever the true distribution \(G\) is, its nonparametric Bayes estimate converges to it, and the choice for the distribution \(G_0\) is not critical. We take for \(G_0\) a normal distribution with mean \(\mu_g\) and variance \(\sigma_g^2\). The Dirichlet process prior for \(G\) is a probability distribution on the space of all possible distributions for the unobserved instrument, and \(G_0\) can be seen as the location parameter (see also appendix 7A). The parameter \(\alpha\) acts as a precision parameter (Escobar, 1994, Escobar and West, 1998): when \(\alpha\) is very large, the Dirichlet process prior \(G\) for \(\theta_{ij}\) is very close to the baseline prior, and when \(\alpha\) is small, \(G\) is not and is likely to concentrate on a few distinct atoms. The expected number of distinct values of \(\theta_{ij}\) is approximately equal to \(\alpha \log[(n + \alpha)/\alpha]\) (Antoniak, 1974, or Escobar, 1994). This number, however, is unknown and estimated from the data.
Throughout the following we assume that the errors are normally distributed. Furthermore, let \( z_{ij} = (y_{ij}, x_{ij})' \), \( \mu_{z,ij} = (\tilde{x}_{ij}\beta, \tilde{\theta}_{ij})' \), where \( \tilde{x}_{ij} = (1, x_{ij}) \), \( \beta = (\beta_0, \beta_1)' \), and \( s \) stands for 'structural'. The density function (likelihood) for the structural model given in (7.2) is given by

\[
p(z_{ij}|\theta_{ij}, \beta, \Sigma) = (2\pi)^{-1}|\Sigma|^{-1/2} \times \exp\left(-\frac{1}{2}(z_{ij} - \mu_{z,ij})'\Sigma^{-1}(z_{ij} - \mu_{z,ij})\right),
\]

which can also be written in the reduced form as

\[
p(z_{ij}|\theta_{ij}, \beta, \Sigma) = (2\pi)^{-1}|\Omega|^{-1/2} \times \exp\left(-\frac{1}{2}(z_{ij} - \mu_{z,ij}')\Omega^{-1}(z_{ij} - \mu_{z,ij}')\right),
\]

where \( \mu_{z,ij} = (\beta_0 + \beta_1\tilde{\theta}_{ij}, \tilde{\theta}_{ij})' \) and \( \Omega = B\Sigma B' \) with

\[
B = \begin{bmatrix} 1 & \beta_1 \\ 0 & 1 \end{bmatrix}.
\]

In the next subsection we discuss the Dirichlet Process prior for \( \theta_{ij} \) in more detail.

### 7.2.1 The Dirichlet process prior for \( \theta_{ij} \)

Conditional on \( \beta, \Sigma, G_0, \alpha, \) and \( z_{ij} = (y_{ij}, x_{ij}) \), the Dirichlet process prior for \( \theta_{ij} \) implies the following conditional posterior distribution (see e.g. Escobar, 1994, Escobar and West, 1998)

\[
p(\theta_{ij}|\theta_{-ij}, \beta, \Sigma, \alpha, G_0, z_{ij}) \sim q_{0(ij)}h(\theta_{ij}|\beta, \Sigma, G_0, z_{ij}) + \\
+ \sum_{l=1}^{n} \sum_{k=1}^{m_l} q_{l(k|ij)}\delta_{\theta_k}(\theta_{ij}),
\]

(7.8)
where \( \theta_{ij}^\ast \) denotes the \((n_t - 1) \times 1\) vector obtained from \( \theta = (\theta_{11}, \ldots, \theta_{mn})' \) with the \( i,j \)-th element deleted, and \( \delta_{0k} (\theta_{ij}) = 1 \) if \( \theta_{ij} = \theta_{lk} \) and zero otherwise. Furthermore,

\begin{itemize}
  \item \( h(\theta_{ij} | \beta, \Sigma, G_0, z_{ij}) \propto p(z_{ij} | \theta_{ij}, \beta, \Sigma) g_0(\theta_{ij} | \mu_g, \sigma_g^2) \) is the density of the ‘baseline’ posterior distribution for \( \theta_{ij} \);
  \item \( q_{0(ij)} \propto \alpha \int p(z_{ij} | \theta_{ij}, \beta, \Sigma) g_0(\theta_{ij} | \mu_g, \sigma_g^2) d\theta_{ij} \), i.e. \( q_{0(ij)} \) is proportional to \( \alpha \) times the marginal distribution of \( z_{ij} \) where \( \theta_{ij} \) is integrated out under the baseline prior \( G_0 \);
  \item \( q_{lk(ij)} \propto p(z_{ij} | \beta, \Sigma, \theta_{lk}) \), i.e. \( q_{lk(ij)} \) is proportional to the density of \( z_{ij} \) conditional on \( \theta_{ij} = \theta_{lk} \);
  \item and \( q_{0(ij)} + \sum_{l \neq k}^n \sum_{k=1}^m q_{lk(ij)} = 1 \), i.e. \( q_{0(ij)} \) and \( q_{lk(ij)} \) are normalized to 1.
\end{itemize}

The intuition behind the above scheme is as follows. If observation \( ij \) has a relatively large (small) residual using observation \( lk \)’s value \( \theta_{lk} \), then it is relatively less (more) likely that observation \( ij \)’s value for the latent instrument is chosen as \( \theta_{lk} \). Furthermore, the smaller the residual for observation \( ij \), while assuming its value for the latent instrument is \( \mu_g \), the greater the probability that observation \( ij \) gets a new value for its latent instrument. In fact, at this point \( G \) has been integrated out and it is conceptually easy to sample from the above distribution using the following scheme:

\[
\theta_{ij} | \theta_{ij}^\ast, \beta, \Sigma, \alpha, G_0, z_{ij} \begin{cases} = \theta_{lk} & \text{with probability } q_{lk(ij)} \\ \sim h(\theta_{ij} | \beta, \Sigma, G_0, z_{ij}) & \text{with probability } q_{0(ij)}. \end{cases}
\] (7.9)

The ease of implementation, however, depends on whether the likelihood is easy to evaluate, whether the density \( h(\theta_{ij} | \beta, \Sigma, G_0, z_{ij}) \) is of manageable form, and whether \( q_{0(ij)} \) can be easily computed. When \( G_0 \) is a conjugate prior, which is often not a strong assumption, the marginal distribution of \( z_{ij} \) is known analytically for the computation of \( q_{0(ij)} \). Otherwise it may be possible to compute it numerically, or to apply methods that circumvent the computation of \( q_{0} \), see Escobar and West (1998).
Chapter 7  A Nonparametric Bayesian LIV approach

Antoniak (1974) shows that clustering of the \( \theta_{ij} \)'s is inherent to the Dirichlet process. I.e., the values \( \theta_{ij} \) typically reduce to \( \bar{n} < n_t \) distinct values or clusters. These values are drawn from \( G_0 \), which can be seen using the Polya Urn\(^1\) representation of the Dirichlet process (Escobar, 1994 or MacEachern, 1998). It is this description of the vector \( \theta \) that leads to the view of the mixture of Dirichlet process model as a mixture model (cf. MacEachern, 1998).

We denote these \( \bar{n} \) distinct values of \( \theta_{ij} \) by \( \tilde{\theta}_s \), \( s = 1, \ldots, \bar{n} \). In the following, the superscript “−” denotes that observation \( ij \) is left out of the conditioning, i.e. \( n^-_s \leq n_t - 1 \) is the number of observations in ‘cluster’ \( s \) with common value \( \tilde{\theta}_s \) without observation \( ij \), and \( \bar{n}^- \) is the number of distinct clusters when observation \( ij \) is removed (see also MacEachern, 1998). Now the conditional posterior distribution of \( \theta_{ij} \) is given by

\[
p(\theta_{ij}|\bar{\theta}_{ij}, \beta, \Sigma, \alpha, G_0, z_{ij}) \sim q_{0(ij)}(\theta_{ij}|\beta, \Sigma, G_0, z_{ij}) + \sum_{l=1}^{\bar{n}^-} n^-_l \tilde{q}_{l(ij)} h(\theta_{ij}|\beta, \Sigma, \tilde{\theta}_l) \delta_{\tilde{\theta}_l}(\theta_{ij}),
\]

(7.10)

where \( \tilde{q}_{l(ij)} \) is proportional to the likelihood \( p(z_{ij}|\beta, \Sigma, \tilde{\theta}_l) \) given that observation \( ij \) is in cluster \( l \) and \( \delta_{\tilde{\theta}_l}(\theta_{ij}) = 1 \) if \( \theta_{ij} = \tilde{\theta}_l \) and 0 otherwise. The other elements are defined as before. From (7.10) it can be seen that the full conditional posterior distribution of \( \theta_{ij} \) is a mixture of a continuous distribution and a discrete distribution with weights on the distinct values for the latent instrument (excluding \( \theta_{ij} \)).

We assume that the baseline distribution \( G_0 \) of the Dirichlet process is a univariate normal distribution with unknown mean \( \mu_g \) and variance \( \sigma_g^2 \). We take conjugate priors for both parameters, i.e. a normal distribution for \( \mu_g \) and an inverse gamma distribution for \( \sigma_g^2 \). The parameter \( \alpha \) (the dispersion parameter for the Dirichlet process) is an important parameter since it determines how

\(^1\)Simply stated, the Polya urn problem concerns an urn that contains (say) \( c_i \) balls of color \( i \). Each time, a ball is randomly taken from the urn and the probability that it has color \( i \) is proportional to the number of balls of that color in the urn. Then it is replaced along with another ball of the same color, etc..
A simple multilevel model with a general latent instrument

‘close’ the unknown distribution \( G \) gets to \( G_0 \). If this value is unknown, as for most applications, a prior distribution \( p(\alpha) \) can be specified and the data is used to learn about this parameter, see Escobar and West (1995), West (1992), Escobar (1994). We follow West (1995) and specify a gamma distribution as prior for \( \alpha \). Finally, for \( \beta \) and \( \Sigma \) we take a normal distribution and an inverted two-dimensional Wishart distribution, respectively, as priors. The parameters of the prior distributions are known and are specified such that they reflect ‘vague’ or little information.

The complete nonparametric Bayes LIV model is rather complex and it is not possible to derive closed form expressions for the joint and marginal posterior distributions of the parameters \( \beta \) and \( \Sigma \), and the other parameters \( \mu_g, \sigma_g^2, \theta, \) and \( \alpha \). However, Escobar’s (1994) MCMC results can be used to approximate the nonparametric Bayes model, where the full conditional distribution of \( \theta_{ij} \)’s is a mixture of a discrete distribution with weights on the other \( \theta_{ik} \)’s, \( lk \neq ij \) and \( G_0 \), see (7.8) and (7.10). His work has been modified and extended in several ways and the resulting chains are relatively easy to implement and have been empirically shown to move quickly through the parameter space (Dey, Müller and Sinha, 1998). The full conditional distributions of the other parameters can be derived straightforwardly when the priors are conjugate. We discuss this in more detail next.

7.2.2 MCMC estimation

Depending on the form of the full conditional we use either (7.5) or (7.6) for the likelihood. The likelihood of the complete sample is computed as \( \prod_{i=1}^{n} \prod_{j=1}^{m_i} p(z_{ij} | \theta_{ij}, \beta, \Sigma) \). We use the following expression for \( \Sigma^{-1} \):

\[
\Sigma^{-1} = \begin{bmatrix}
\frac{\sigma_e^2}{|\Sigma|} & -\frac{\sigma_{e\mu}}{|\Sigma|} \\
-\frac{\sigma_{e\mu}}{|\Sigma|} & \frac{\sigma_{\mu\mu}}{|\Sigma|}
\end{bmatrix} = \begin{bmatrix}
\sigma^{(11)} & \sigma^{(12)} \\
\sigma^{(21)} & \sigma^{(22)}
\end{bmatrix},
\]

(7.11)

where \( \sigma^{(12)} = \sigma^{(21)} \), and \( |\Sigma| = \sigma_e^2 \sigma_{\mu\mu} - (\sigma_{e\mu})^2 \). The joint posterior distribution for the parameters \( (\beta, \Sigma, \mu_g, \sigma_g^2, \alpha) \) is proportional to the likelihood times the priors, i.e.
Chapter 7 A Nonparametric Bayesian LIV approach

\[ p(\beta, \Sigma, \mu_g, \sigma_g^2, \alpha) \propto \left( \prod_{i=1}^{n} \prod_{j=1}^{m_i} p(z_{ij}|\theta_{ij}, \beta, \Sigma) \right) \times p(G|G_0, \alpha) \times \]
\[ \times p(\alpha) \times p(\mu_0) \times p(\sigma_g^2) \times p(\beta), \quad (7.12) \]

where \( p(z_{ij}|\theta_{ij}, \beta, \Sigma) \) is given in (7.5) or (7.6). This joint posterior density is intractable analytically but Markov Chain Monte Carlo (MCMC) methods can be used to generate random draws indirectly without having to calculate the joint posterior explicitly. The MCMC chain is implemented via the following full conditional distributions of the joint posterior distribution:

1. \( p(\Sigma|\beta, \alpha, \mu_g, \sigma_g^2, \theta, z) \),
2. \( p(\beta|\Sigma, \alpha, \mu_g, \sigma_g^2, \theta, z) \),
3. \( p(\mu_g|\beta, \Sigma, \alpha, \sigma_g^2, \theta, z) \),
4. \( p(\sigma_g^2|\beta, \Sigma, \alpha, \mu_g, \theta, z) \), and
5. \( p(\theta_{ij}|\theta_{ij}, \beta, \Sigma, \alpha, \mu_g, z) \) for each \( i = 1, ..., n, j = 1, ..., m_i \),

where \( \theta \) and \( z \) are the \( n \times 1 \) vectors containing the elements \( \theta_{ij} \) and \( z_{ij} \). When the Markov chain stabilizes on a (relatively) small number of distinct values \( \tilde{\theta}_s, s = 1, ..., \tilde{n} \), it is unlikely that new values for \( \theta \) are generated and, hence, the chain gets ‘stuck’ and has undesirable mixing properties. In order to prevent the chain from getting stuck on a few nodes, West, Müller and Escobar (1994) propose to ‘remix’ \( \tilde{\theta}_s, s = 1, ..., \tilde{n} \) after each iteration of the MCMC algorithm (see also Escobar and West, 1998). Let \( S = (S_{11}, ..., S_{nmn}) \) denote the cluster structure, that is \( S_{ij} = s \) if \( \theta_{ij} = \tilde{\theta}_s \) for \( i = 1, ..., n, j = 1, ..., m_i \) and \( s = 1, ..., \tilde{n} \). Given this configuration, the full conditional distribution \( p(\tilde{\theta}_s|S, \tilde{n}, \beta, \Sigma, \mu_g, \sigma_g^2, z), s = 1, ..., \tilde{n} \), can be used to generate a new set of values \( \tilde{\theta} \) to provide more movement in the chain which facilitates convergence. As we show later on, the remixing step typically involves drawing \( \tilde{n} \) values for \( \tilde{\theta} \) from a distribution that has a density somewhat similar to \( h \) in (7.10).
In the following we provide the specification of the full conditional distributions above. More detailed information is given in appendix 7B.

**Full conditional distribution of $\Sigma$.** The full conditional distribution for $\Sigma$ reduces to

$$p(\Sigma|\beta, \theta, z) \propto \left( \prod_{i=1}^{n} \prod_{j=1}^{m_i} p(z_{ij}|\theta_{ij}, \beta, \Sigma) \right) \times p(\Sigma|\omega, \Psi), \tag{7.13}$$

from which it follows that $\Sigma$ is sampled from a two-dimensional inverse Wishart distribution with parameters $\omega + \tilde{n}$ and $(\sum_{i} \sum_{j} (z_{ij} - \mu_{ij}^z)(z_{ij} - \mu_{ij}^z)^T + \Psi)$.

**Full conditional distribution of $\beta$.** Similarly, the full conditional distribution of the regression parameters $\beta$ is obtained from combining the likelihood and a normal prior distribution, i.e.

$$p(\beta|\Sigma, \theta, z) \propto \left( \prod_{i=1}^{n} \prod_{j=1}^{m_i} p(z_{ij}|\theta_{ij}, \beta, \Sigma) \right) \times p(\beta|\mu_\beta, \Sigma_\beta), \tag{7.14}$$

Hence, $\beta$ is sampled from a normal distribution with mean

$$C^{-1} \left( \sum_{i,j} \tilde{x}_{ij}' (\sigma^{(1)} y_{ij} + \sigma^{(2)} (x_{ij} - \theta_{ij})) + \Sigma_\beta^{-1} \mu_\beta \right),$$

and variance-covariance $C^{-1}$, where $C = \sigma^{(1)} \sum_{i,j} \tilde{x}_{ij}' \tilde{x}_{ij} + \Sigma_\beta^{-1}$ and $\tilde{x}_{ij} = (1, x_{ij})$.

**Full conditional of $\theta_{ij}$.** A sample for each $\theta_{ij}$ can be obtained using (7.10). Hence,

1. sample a proposed ‘cluster’ value $c_{ij}$ from the integers $\{0, 1, \ldots, \tilde{n}-\}$
   with probabilities proportional to $\{q_{0(ij)}, n^{-1} q_{1(ij)}, \ldots, n^{-1} q_{\tilde{n}-(ij)}\}$. 
2. If \( c_{ij} \in \{1, ..., \tilde{n}^-\} \), then set \( \theta_{ij} = \tilde{\theta}_{c_{ij}} \), and if \( c_{ij} = 0 \), then draw a new value \( \theta_{ij} \) from \( h(\theta_{ij}|\beta, \Sigma, G_0, z_{ij}) \), which is the density of a univariate normal distribution with mean \( \tilde{C}^{-1}(\sigma^{(22)}x_{ij} + \sigma^{(12)}(y_{ij} - \bar{x}_j\beta) + \mu_g/\sigma_g^2) \) and variance \( \tilde{C}^{-1} = (\sigma^{(22)} + 1/\sigma_g^2)^{-1} \).

We provide more detail on the form of the probabilities \( q_{0(ij)}, q_{1(ij)}, ..., q_{\tilde{n}-(ij)} \) in appendix 7B.

**Remix \( \tilde{\theta} \).** Let \( J_l \) denote the set of observations for which \( \theta_{ij} = \tilde{\theta}_l, l = 1, ..., \tilde{n} \).

The full conditional of \( \tilde{\theta}_l \) is proportional to

\[
\left\{ \prod_{i,j \in J_l} p(z_{ij}|\tilde{\theta}_l, \beta, \Sigma) \right\} g_0(\tilde{\theta}_l|\mu_g, \sigma_g^2),
\]

for \( l = 1, ..., \tilde{n} \), see e.g. West, Müller and Escobar (1994), and Escobar and West (1998). The derivation of this distribution is more or less similar to the derivation of \( h(\theta_{ij}|\beta, \Sigma, G_0, z_{ij}) \) above.

The \( \tilde{\theta}_l \)'s are updated by replacing the current values by new values that are drawn from an univariate normal distribution with mean \( \tilde{C}^{-1}\left(\sum_{i,j \in J_l} x_{ij} + \sigma^{(12)}\sum_{i,j \in J_l} (y_{ij} - \bar{x}_j \beta) + \mu_g/\sigma_g^2\right) \) and variance \( \tilde{C}^{-1} = \sigma^{(22)} + 1/\sigma_g^2 \).

**Full conditional of \( \mu_g \).** The parameters \( \tilde{\theta}_l \) are independent and identically distributed from \( G_0(.|\mu_g, \sigma_g^2) \) and \( \mu_g \) enters the model only through \( G_0 \). Hence, the full conditional distribution for \( \mu_g \) is given by (West, Müller and Escobar, 1994)

\[
p(\mu_g|\sigma_g^2, \tilde{\theta}, \tilde{n}, z) \propto \left\{ \prod_{l=1}^{\tilde{n}} g_0(\tilde{\theta}_l|\mu_g, \sigma_g^2) \right\} p(\mu_g|\mu_0, \sigma_0^2), \quad (7.15)
\]

where both densities are normal densities. Hence, a new value for \( \mu_g \) is drawn from a normal distribution with mean \( C^{-1}(\sum_{l} \tilde{\theta}_l/\sigma_g^2 + \mu_0/\sigma_0^2) \) and variance \( C^{-1} \), with \( C = (1/\sigma_0^2 + \tilde{n}/\sigma_g^2) \).
Full conditional of $\sigma^2_g$. Similarly, the full conditional of $\sigma^2_g$ reduces to

$$p(\sigma^2_g | \mu_g, \tilde{\theta}, \tilde{n}, z) \propto \left\{ \prod_{l=1}^{\hat{n}} g_0(\tilde{\theta}_l | \mu_g, \sigma^2_g) \right\} p(\sigma^2_g | c, d), \quad (7.16)$$

where the prior for $\sigma^2_g$ is an inverted gamma distribution. It follows that $\sigma^2_g$ is sampled from an inverted gamma distribution with parameters $c + \tilde{n}/2$ and $\frac{1}{2} \sum_j (\tilde{\theta}_j - \mu_g)^2 + d$.

Full conditional for $\alpha$. The full conditional distribution of $\alpha$, the ‘dispersion’ of the Dirichlet process, reduces to $p(\alpha | \tilde{n}, n_t)$. In fact, when the prior $p(\alpha)$ is a gamma density with parameters $\tau_\alpha, \gamma_\alpha$, it is possible to obtain an exact expression for the full conditional of $\alpha$. Escobar and West (1995), or West (1992) show that a new value for $\alpha$ can be obtained in two steps:

1. sample an auxiliary value $\eta$ from $p(\eta | \alpha, \tilde{n}, n_t) \sim \text{Beta}(\alpha + 1, n_t)$, i.e. a beta distribution with mean $\frac{\alpha + 1}{\alpha + n_t + 1}$.

2. Then, sample $\alpha$ from the following mixture of gamma’s: $p(\alpha | \eta, \tilde{n}, n_t) \sim \pi_n \text{Gamma}[\tau_\alpha + \tilde{n}, \gamma_\alpha - \log(\eta)] + (1 - \pi_n) \text{Gamma}[\tau_\alpha + \tilde{n} - 1, \gamma_\alpha - \log(\eta)]$, where $\frac{\pi_n}{1 - \pi_n} = \frac{\tau_\alpha + \tilde{n} - 1}{\tilde{n} (\gamma_\alpha - \log(\eta))}$.

This completes the specification of the MCMC chain. Escobar (1994) and Escobar and West (1995) prove convergence theorems for MCMC chains that use a Dirichlet process prior. Using suitable starting values, the above scheme can be iterated many times to obtain a sample of any size from the true posterior distribution. An important question is to determine how often this scheme has to be repeated to ensure convergence of the chain, see for instance Cowles and Carlin (1996), or Brooks and Roberts (1998). We present simulation results for this model in section 7.4. In the following we consider a more general hierarchical regression model.
7.3 **Endogenous subject-level covariates and random coefficients**

In this section we consider general random coefficients models (e.g. Lenk et al., 1996) with possible endogenous subject-level covariates, in which case parameter estimates using standard estimation techniques are no longer guaranteed to be unbiased. This may occur, for instance, when relevant covariates that are correlated with included covariates, are omitted, or when some of the covariates are measured with error. For instance, the self-reported measure for knowledge about the microcomputer market to explain part of the variability of the random regression coefficients, used by Lenk et al. (1996), is possibly measured with error because individuals may find it difficult to adequately express their knowledge by a few statements. Besides, if the measures used are not, or only partly, related with the constructs that are actually searched for, the observed data that is used in estimation contains measurement error. As far as we know there are no other studies that consider regressor-error dependencies at this stage of the model, but as will become clear the estimated model parameters may be biased in presence of such endogenous covariates using standard estimation techniques. To be more specific, here we investigate a more general form of level-2 dependencies as in section 6.1 (see also the discussion about random coefficient models in section 6.6).

The model we consider is a standard linear two-level model with random coefficients. We assume that a set of individual level covariates are available to explain part of the variance of the random coefficients. The model is given by

\[ y_{ij} = x'_{ij} \beta_i + \epsilon_{ij} \]
\[ \beta_i = \gamma_c + \gamma z_i + \eta_i, \]

(7.17)

with \( i = 1, ..., n \) and \( j = 1, ..., m \), or \( m_i \). \( x_{ij} \) is a set of explanatory variables (e.g. a design matrix in conjoint analysis), which is, as \( \beta_i \), a \( k \times 1 \) vector. The individual-level covariates are given by \( z_i = (z'_{i1}, z'_{i2})', \) where \( z_{ij} \) is a \( l_1 \times 1 \) vector of potential endogenous covariates that are correlated with \( \eta_i \) (but not
7.3 Endogenous subject-level covariates and random coefficients

with $\epsilon_{ij}$, and $z_2$ is a $l_2 \times 1$ vector containing the exogenous covariates. $\gamma$ is a $k \times (l_1 + l_2)$ matrix, and the constant $\gamma_c$ is a $k \times 1$ vector. We also write $\gamma_0 = (\gamma', \gamma_c')'$, i.e. the matrix $\gamma_0$ represents the effect of the covariates $z_i$ on the regression coefficients $\beta_i$. The model for $\beta_i$ is a (latent) multivariate regression model, where $E(\eta_i) = 0$ and $\text{Var}(\eta_i) = \Sigma_{\beta\beta}$. We assume that all the $\epsilon_{ij}$'s are independent across and within subjects, with mean 0 and variance $\sigma^2$.

We are not restricted to within subject independency and the results presented here can be generalized to a situation where $\text{Var}(\epsilon_i) = \Sigma$, or $\Sigma_i$.

The nonparametric Bayes LIV approach presented in the previous section can be used in the model for $\beta_i$ to solve for possible biases in the presence of endogenous covariates $z_1$. Here the idea of latent instruments is in particular useful since obtaining valid observed instruments at this stage of the model is highly problematic and ambiguous. The latent instruments are included as follows

$$z_{1i} = \theta_i + \alpha z_{2i} + \xi_i, \quad (7.18)$$

where $\theta_i$ is a $l_1 \times 1$ vector of unobserved instruments, $\alpha$ a $l_1 \times l_2$ matrix that contain the effects of the exogenous covariates on the endogenous covariates$^2$, and $\xi_i$ is a $l_1 \times 1$ vector of errors, which has expectation zero. The dependency between the covariates $z_{1i}$ and $\eta_i$, i.e. the endogeneity, is caused by a nonzero covariance between $\eta_i$ and $\xi_i$. The variance covariance matrix of $(\eta_i, \xi_i)$ is the $(k + l_1) \times (k + l_1)$ matrix $\Lambda$, which is assumed to be positive definite, and contains the block matrices $\Sigma_{\beta\beta}$, $\Sigma_{\beta z_1}$, and $\Sigma_{z_1 z_1}$.

As in the previous section, the ‘latent’ instrument $\theta_i$ has an unknown distribution $G$, which has a Dirichlet process prior with parameters $\rho$ and $G_0$. $G_0$ is the normal density with mean $\mu_\theta$, which is a $(l_1 \times 1)$-vector, and a $(l_1 \times l_1)$ variance-covariance matrix $V_\theta$. In the following we give the full conditionals for the MCMC scheme. We use standard conjugate priors for the parame-

$^2$This parameter $\alpha$ is not to be confused with the ‘dispersion’ parameter $\alpha$ of the Dirichlet process in the previous section. For the model in this section we use $\rho$ for that purpose.
ters which are specified such that they represent no or vague prior knowledge. More details on the derivation of the full conditionals is given in appendix 7C.

7.3.1 Estimating the hierarchical model with general latent instruments

The unknown parameters of the model in (7.17) and (7.18) are: $\beta_i$, $\sigma^2$, $\gamma_0$, $\theta_i$, $\alpha$, $\Lambda$, $\rho$, $\mu_\theta$ and $V_\theta$. The joint posterior distribution, conditionally on the data $X$, $Y$, and $Z$, is formed as a product of the likelihood function, obtained from the first equation of model (7.17), and the prior densities. Since no closed-form expressions for the joint posterior density and marginal posterior densities are available, we use MCMC sampling to approximate a sample from the true marginal posterior densities. Assuming normal distributions for the error terms, the full conditional distributions for the parameters with conjugate priors can be obtained similar to section 7.2. Here we generally use multivariate distributions to accommodate for having a vector $\beta_i$ and possible several endogenous variables $z_i$. See Escobar (1994) and Dey, Müller and Sinha (1998) for more details on general multivariate nonparametric Bayesian estimation. The MCMC scheme is completed by iterating the following conditional distributions:

1. $p(\sigma^2 | \beta_i, i = 1, ..., n; Y, X)$;
2. $p(\beta_i | \sigma^2, \gamma_0, \Lambda, \theta_i, \alpha; Y, X, Z)$, for $i = 1, ..., n$;
3. $p(\Lambda | \beta_i, \theta_i, i = 1, ..., n, \gamma_0, \alpha; Z)$;
4. $p(\gamma_0 | \beta_i, \theta_i, i = 1, ..., n, \Lambda, \alpha; Z)$;
5. $p(\alpha | \beta_i, \theta_i, i = 1, ..., n, \Lambda, \gamma_0; Z)$;
6. $p(\theta_i | \theta_i^-, \beta_i, \Lambda, \gamma_0, \alpha; Z)$, where $\theta_i^- = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n)$, for $i = 1, ..., n$;
7. $p(\mu_\theta | \tilde{\theta}_i, l = 1, ..., \tilde{n}, V_\theta)$, where $\tilde{\theta}_i$ are the $\tilde{n} \leq n$ distinct values for $\theta_i$ that generally arise from the clustering structure in the Dirichlet process;
8. $p(V_\theta | \tilde{\theta}_i, l = 1, ..., \tilde{n}, \mu_\theta)$, and
9. \( p(\rho|\tilde{n}, n) \),

and, as before, a remixing step for the different values of \( \tilde{\theta}_l, l = 1, \ldots, \tilde{n} \). In the following we give the specific distributions, see appendix 7C for more details.

**Full conditional distribution of \( \sigma^2 \).** The full conditional for \( \sigma^2 \) is proportional to the likelihood times the inverse gamma prior distribution for \( \sigma^2 \) with parameters \( \tau_0 \) and \( \eta_0 \). Hence, a new value for \( \sigma^2 \) is drawn from an inverse gamma distribution with parameters \( \tau_0 + nm/2 \) and \( (1/2) \sum_{i,j} (y_{ij} - x_{ij}'\beta_i)^2 + \eta_0 \).

**Full conditional distribution of \( \beta_i \).** The full conditional for \( \beta_i \) can be obtained from

\[
p(\beta_i|\text{rest, data}) \propto \left\{ \prod_{j=1}^{m} p_1(y_{ij}|\beta_i, \sigma^2) \right\} p_k(\beta_i|z_{1i}, z_{2i}, y_0, \Lambda, \theta, \alpha) \tag{7.19}
\]

where \( p_1 \) denotes the \( x \)-variate normal density. The latter conditional distribution in (7.19) is obtained from the joint distribution \( p_{k+l_1}(h_i|z_{2i}, \gamma, \Lambda, \theta, \alpha) \), where \( h_i = (\beta'_i, z'_{1i})' \) is a \((k+l_1) \times 1 \) vector. Hence, \( p_k(\beta_i|z_{1i}, z_{2i}, y_0, \Lambda, \theta, \alpha) \) is a \( k \)-variate normal distribution with mean

\[
\mu_{\beta_{z_1}} = \gamma_c + \gamma z_i + \Sigma_{\beta_{z_1}}^{-1} \Sigma_{z_{1z_1}}^{-1} (z_{1i} - (\theta_i + \alpha z_{2i})) \tag{7.20}
\]

and variance-covariance

\[
\Sigma_{\beta_{z_1}} = \Sigma_{\beta\beta} - \Sigma_{\beta z_1} \Sigma_{z_{1z1}}^{-1} \Sigma_{\beta' z_1} \tag{7.21}
\]

(e.g. theorem 3.6, Greene, 2000). Let \( \tilde{C} = \frac{1}{\sigma^2} \sum_j x_{ij}'y_{ij} + \Sigma_{\beta\beta,z_1}^{-1} \). It follows that the full conditional for \( \beta_i \) is a \( k \)-variate normal density with mean \( \tilde{\mu}_{\beta_{z_1}} = \left( \gamma_c + \gamma z_i, (\theta_i + \alpha z_{2i}) \right)' \) and variance-covariance \( \tilde{C}^{-1} \).

**Full conditional distribution of \( \Lambda \).** As before, let \( h_i = (\beta'_i, z'_{1i})' \) and define \( \mu_{h_i} = ((\gamma_c + \gamma z_i)', (\theta_i + \alpha z_{2i})') \). The full conditional for \( \Lambda \) is obtained as
Chapter 7 A Nonparametric Bayesian LIV approach

the product of the joint density of $h_i$ across $i = 1, ..., n$, and the density of the inverted Wishart prior distribution with parameters $(c, D)$. Hence, $\Lambda$ is sampled from an inverted Wishart $k+l_1$-distribution with parameters $n + c$ and

$$\sum_{i=1}^{n} (h_i - \mu_{h_i}) (h_i - \mu_{h_i})' + D.$$ 

**Full conditional for $\gamma_0$.** The full conditional distribution for $\gamma_0$ can be obtained by vectorizing the model for $(\beta_i, z_{1i})$. I.e. we stack the rows of $\beta_i$ and $z_{1i}$ as follows (see e.g. Lenk, 2001)

$$\tilde{\beta}_i = \text{vec}(\beta_i) = \text{vec}(\gamma_c) + \text{vec}(\gamma_{z_i}) + \text{vec}(\eta_i)$$

$$\tilde{\beta}_i = (z_{0i}' \otimes I_k) \tilde{\gamma}_0 + \tilde{\eta}_i,$$  \hspace{1cm} (7.22)

with $z_{0i} = (z_i', 1)'$, a $(l + 1) \times 1$-vector, and

$$\tilde{z}_{1i} = \text{vec}(z_{1i}) = \text{vec}(\theta_i) + \text{vec}(\alpha z_{2i}) + \text{vec}(\xi_i)$$

$$\tilde{z}_{1i} = \tilde{\theta}_i + (z_{2i}' \otimes I_{l_1}) \tilde{\alpha} + \tilde{\xi}_i.$$  \hspace{1cm} (7.23)

We write $\tilde{z}_{0i} = z_{0i}' \otimes I_k$ and $\tilde{z}_{2i} = z_{2i}' \otimes I_{l_1}$. Furthermore, let $\tilde{z}_{10i} = \tilde{z}_{1i} - \tilde{\theta}_i - \tilde{z}_{2i} \tilde{\alpha}$, $\Lambda = \text{var}(\tilde{\eta}_i, \tilde{\xi}_i)$ and\footnote{See Greene (2000), formula (2-74), for a general expression of the inverse of a $2 \times 2$ partitioned matrix.}

$$\Lambda^{-1} = \begin{bmatrix} \Lambda^{(11)} & \Lambda^{(12)} \\ \Lambda^{(21)} & \Lambda^{(22)} \end{bmatrix},$$  \hspace{1cm} (7.24)

where $\Lambda^{(12)} = \Lambda^{(21)}'$, $\Lambda^{(11)}$ is $k \times k$, $\Lambda^{(12)}$ is $k \times l_1$, and $\Lambda^{(22)}$ is a $l_1 \times l_1$ matrix. The vectorized system for $\tilde{\beta}_i$ and $\tilde{z}_{1i}$ has a 'standard' multivariate normal form. It follows from appendix 7C that the values for $\tilde{\gamma}_0$ are drawn from a $(l_1 + l_2 + 1)$-variate normal distribution with mean

$$\tilde{C}^{-1} \left[ \sum_i \tilde{z}_{0i} \left( \Lambda^{(12)} \tilde{z}_{10i} + \Lambda^{(11)} \tilde{\beta}_i \right) + V_y^{-1} m_y \right].$$  \hspace{1cm} (7.25)
7.3 Endogenous subject-level covariates and random coefficients

and variance-covariance \( \tilde{C}^{-1} \), where \( \tilde{C} = \sum_i \tilde{z}_i' \Lambda^{(11)} \tilde{z}_i + V^{-1} \).

**Full conditional for \( \alpha \).** Using similar arguments as for the matrix of regression coefficients \( \gamma \), the full conditional distribution for \( \alpha \) can be easily obtained after vectorization. I.e., let

\[
\tilde{\beta}_0 = \bar{\eta}_i, \\
\tilde{z}_{10} = \tilde{z}_2 \alpha + \bar{\xi}_i, \tag{7.26}
\]

where \( \tilde{\beta}_0 = \text{vec}(\beta_i - \gamma_c - \gamma z_i) \), \( \tilde{z}_{10} = \text{vec}(z_{1i} - \theta_i) \), and \( \tilde{z}_2 = z'_2 \otimes I_{l_i} \). It follows that the full conditional density for \( \tilde{\alpha} \) is from a multivariate normal distribution with mean (let \( \tilde{C} = \sum_i \tilde{z}_i' \Lambda^{(22)} \tilde{z}_i + V^{-1} \))

\[
\tilde{C}^{-1} \left[ \sum_i \tilde{z}_2' \left( \Lambda^{(21)} \tilde{\beta}_0 + \Lambda^{(22)} \tilde{z}_{10} \right) + V^{-1} m_\alpha \right], \tag{7.27}
\]

and variance \( \tilde{C}^{-1} \).

**Full conditional for \( \theta_j \).** The structure of the full conditional distributions for each of the unobserved instruments \( \theta_j , i = 1, ..., n \), is derived in a similar way as above for the simple linear multilevel model. The full conditional distribution for \( \theta_j \) has the following form, with \( \theta_i^- = (\theta_1, ..., \theta_i-1, \theta_{i+1}, ..., \theta_n) \),

\[
[\theta_i^- | \theta_i, \text{rest, data}] \sim q_{0(i)} h(\theta_i | \text{rest, data}) + \sum_{l=1}^{\tilde{n}^-} n_l^- q_{(i)} \delta_l(\theta_i^-), \tag{7.28}
\]

where \( \tilde{n}^- \) are the number of different \( \theta_j \)'s, \( j = 1, ..., n \), \( j \neq i \), and \( n_l^- \) are the number of observations in cluster \( l \) when the \( i \)-th observation is removed. Hence,

1. sample a proposed ‘cluster’ value \( c_i \) from the integers \( \{0, 1, ..., \tilde{n}^-\} \) with probabilities proportional to \( \{q_{0(i)}, n_l^- q_{1(i)}, ..., n_l^- q_{\tilde{n}^-}\} \).
2. If $c_i \in \{1, ..., \tilde{n}^\sim\}$, set $\theta_i = \tilde{\theta}_{c_i}$, and if $c_i = 0$, then draw a new value $\theta_i$ from $h(\theta_i | \beta_i, \alpha, \gamma_0, \Lambda, \mu_{\theta}, V_{\theta}; Z)$, which is a multivariate normal distribution with mean $\tilde{C}^{-1}(\Lambda^{(21)} \beta_{0i} + \Lambda^{(22)} z_{10i} + V_{\theta}^{-1} \mu_{\theta})$ and variance $	ilde{C}^{-1} = (\Lambda^{(22)} + V_{\theta}^{-1})^{-1}$, where $\beta_{0i} = \beta_i - \gamma_c - \gamma z_i$ and $z_{10i} = \zeta_{1i} - \alpha z_j$.

In the appendix we give more details on how to compute the probabilities $\{q_{0(i)}, n_1^{-} q_{1(i)}, ..., n_{\tilde{n}^\sim}^{-} q_{\tilde{n}^\sim-(i)}\}$. 

**Remix $\tilde{\theta}$.** The remixing density for $\tilde{\theta}_l$, $l = 1, ..., \tilde{n}$, is proportional to

$$
\left\{ \prod_{i \in J_j} p(h_i | \tilde{\theta}_l, \alpha, \gamma_0, \Lambda; Z) \right\} g_0(\tilde{\theta}_l | \mu_{\theta}, V_{\theta}),
$$

where $J_j$ is the set of indicators of observations belonging to group $j$. This distribution can be obtained in a similar manner as $h(\theta_i | \text{rest, data})$ in the previous subsection. Hence, the remixing density for $\tilde{\theta}_l$ is a multivariate normal distribution with mean

$$
\tilde{C}^{-1}\left(\Lambda^{(21)} \sum_{i \in J_j} \beta_{0i} + \Lambda^{(22)} \sum_{i \in J_j} z_{10i} + V_{\theta}^{-1} \mu_{\theta}\right),
$$

and variance $\tilde{C}^{-1}$, with $\tilde{C} = \tilde{n} \Lambda^{(22)} + V_{\theta}^{-1}$, and where $\beta_{0i}$ and $z_{10i}$ are defined as before.

**Full conditional for $\mu_{\theta}$.** The full conditional distribution for $\mu_{\theta}$ is proportional to

$$
\left\{ \prod_{l=1}^{\tilde{n}} g_0(\tilde{\theta}_l | \mu_{\theta}, V_{\theta}) \right\} p(\mu_{\theta} | m_{\mu}, V_{\mu}),
$$

which are both densities of a multivariate normal distribution. Let $\tilde{C} = \tilde{n} V_{\theta}^{-1} + V_{\mu}^{-1}$. Hence, $\mu_{\theta}$ is sampled from a multivariate normal distribution with mean

$$
\tilde{C}^{-1}\left(V_{\theta}^{-1} \sum_{l=1}^{\tilde{n}} \tilde{\theta}_l + V_{\mu}^{-1} m_{\mu}\right),
$$
and variance $\tilde{C}^{-1}$.

**Full conditional distribution for $V_\theta$.** The prior distribution for $V_\theta$ is an inverted Wishart with parameters $(\tau_v, \Upsilon_v)$. Its full conditional distribution is obtained in a similar way as for $\mu_\theta$ and can be shown to be equal to an inverted Wishart distribution of dimension $l_1$ with parameters $\tau_v + \tilde{n}$ and

$$\sum_{l=1}^{\tilde{n}} \left( \tilde{\theta}_l - \mu_\theta \right) \left( \tilde{\theta}_l - \mu_\theta \right)' + \Upsilon_v.$$

**Full conditional distribution for $\rho$.** The full conditional for $\rho$ is derived in a similar way as the dispersion parameter in the simple multilevel model in the previous section. Assuming a gamma prior distribution with parameters $(\tau_\rho, \gamma_\rho)$, we obtain the following scheme for generating a new value for $\rho$:

1. sample an auxiliary value $\eta$ from $p(\eta|\rho, \tilde{n}, n) \sim \text{Beta}(\rho + 1, n)$, i.e. a beta distribution with mean $\frac{\rho + 1}{\rho + n + 1}$.

2. Then, sample $\rho$ from the following mixture of gamma’s: $p(\rho|\eta, \tilde{n}, n) \sim \pi_n \Gamma[(\tau_\rho + \tilde{n}) + (1 - \pi_n)]\Gamma[(\tau_\rho + \tilde{n} - 1) + \gamma_\rho - \log(\eta)]$, where $\pi_n = \frac{\tau_\rho + \tilde{n} - 1}{\pi_\rho - \log(\eta)}$.

This completes the specification of the MCMC scheme. Similar arguments as for the convergence of the MCMC algorithm for the simple nonparametric Bayes LIV model in subsection 7.2.2 apply. In the following section we illustrate the performance of the proposed two models and estimation schemes using synthetic data.

### 7.4 A simulation study

In this section we discuss the results of two simulation studies to investigate the performance of the models and estimation algorithms proposed in the previous two sections. We first discuss the simulation results for the simple nonparametric Bayes model presented in section 7.2. Here we investigate the performance
of the model for three different choices of the distribution of the latent instrument. Then we present the results for the random coefficients model in section 7.3 and consider a situation with one and two endogenous covariates.

7.4.1 Simulation results for the simple multilevel model

We specified three different distributions for the latent instrument in model (7.2): (1) a discrete distribution with two categories, similar to the bimodal ($\tilde{m} = 2$) case in section 3.5, (2) a continuous gamma distribution, and (3) a $t$ distribution with six degrees of freedom. When the true distribution for $\theta_i$ is discrete, the standard LIV model in chapter 3 is correctly specified (conditionally on knowing the true number of categories of the unobserved instrument) and we expect that the standard LIV model outperforms the nonparametric Bayesian LIV approach. When the unobserved instrument has a $t$ distribution, the model is weakly identified, because it is not identified in case of an exact normal distribution\(^4\).

In all cases the variance of $\theta_{ij}$ is equal to 1.5. Furthermore, we normalized its mean to zero. We took an initial sample size of $n = 1000$ and assumed we had only one observation per individual, i.e. $m_i = 1$ for $i = 1, ..., n$. Furthermore, we took $\beta_0 = 1$, $\beta_1 = 2$, $\sigma^2_{\epsilon} = \sigma^2_{\nu} = 1$, and $\sigma_{\epsilon\nu}$ was taken 0, 0.36, and 0.79, representing a situation with no, moderate and severe endogeneity, respectively. In total, we generated 15 datasets. We discarded the first 5000 iterations of the MCMC chain and saved the final 20000. To reduce the autocorrelation in the MCMC draws, we only used every 10th draw. Convergence was monitored based on iteration plots. We first discuss the results for the main parameters for the bimodal and gamma distributions and compare the results obtained from the nonparametric Bayes LIV method with the classical OLS estimates and the standard LIV estimates. Subsequently, we present our findings when the latent instrument has a $t$ distribution.

\(^4\)We do not have a formal proof of this conjecture. If the distribution of the unobserved instrument is exactly normal, it is identical to the specification of $G_0$ from which the unobserved instruments are drawn. We can expect to end up with either $\tilde{n} = 1$ or $\tilde{n} = n_t$. Both situations are not identifiable. We found support for this using simulated data.
7.4 A simulation study

Results main parameters for bimodal and gamma distribution

Table 7.1: Results main parameters for bimodal distribution for the three situations: \( \sigma_{\epsilon v} = 0 \) (A), \( \sigma_{\epsilon v} = 0.36 \) (B), and \( \sigma_{\epsilon v} = 0.79 \) (C).

<table>
<thead>
<tr>
<th></th>
<th>( \beta_1 )</th>
<th>( \sigma_{\epsilon r}^2 )</th>
<th>( \sigma_{\epsilon v} )</th>
<th>( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>OLS</td>
<td>1.99 (0.020)</td>
<td>1.00 (0.027)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes LIV</td>
<td>2.00 (0.043)</td>
<td>1.01 (0.029)</td>
<td>-0.02 (0.127)</td>
</tr>
<tr>
<td></td>
<td>LIV2</td>
<td>2.00 (0.037)</td>
<td>1.00 (0.028)</td>
<td>-0.02 (0.115)</td>
</tr>
<tr>
<td>B</td>
<td>OLS</td>
<td>2.15 (0.012)</td>
<td>0.95 (0.046)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes LIV</td>
<td>2.01 (0.030)</td>
<td>1.00 (0.050)</td>
<td>0.35 (0.074)</td>
</tr>
<tr>
<td></td>
<td>LIV2</td>
<td>2.00 (0.029)</td>
<td>1.00 (0.050)</td>
<td>0.36 (0.073)</td>
</tr>
<tr>
<td>C</td>
<td>OLS</td>
<td>2.31 (0.015)</td>
<td>0.76 (0.033)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes LIV</td>
<td>2.00 (0.028)</td>
<td>1.01 (0.049)</td>
<td>0.78 (0.030)</td>
</tr>
<tr>
<td></td>
<td>LIV2</td>
<td>2.00 (0.029)</td>
<td>1.00 (0.051)</td>
<td>0.78 (0.031)</td>
</tr>
</tbody>
</table>

The results for the bimodal distribution are presented in table 7.1. We present the mean and standard deviations of the estimated parameters computed across the 15 simulated datasets. For the nonparametric Bayes model we computed the posterior means for \( \beta_1 \), \( \sigma_{\epsilon r}^2 \), \( \sigma_{\epsilon v} \), and \( k \) across the 2000 saved MCMC iterations and, subsequently, we computed the average and standard deviations across these 15 posterior means. We do not report the results for \( \beta_0 \) because it was estimated consistently by OLS since \( x_{ij} \) has mean zero in all cases.

It follows from table 7.1 that the simple nonparametric Bayes model gives approximate unbiased results in all cases. As can be seen, the number of clusters (i.e. different values of \( \theta_{ij} \)), as indicated by \( k \), is a parameter in the nonparametric Bayes model and is also estimated, as opposed to the standard LIV model where \( k \) has to be chosen a priori. The high standard deviations of the estimated values for \( k \) do not mean that it is not estimated precisely, but throughout the MCMC iterations a few high values of \( k \) appear that affect its mean and standard deviation, see for instance figure 7.1. In particular for \( \sigma_{\epsilon v} > 0 \), we find that the number of components estimated by the nonparametric Bayes
LIV model was more close to the true number of components, which was two. It can be seen that the OLS results are biased when the regressor is correlated with the error term, i.e. when $\sigma_{\epsilon \nu} \neq 0$. When $x$ is truly exogenous, OLS is the best alternative and the classical LIV model, which was specified with $k = 2$, is slightly more efficient that the nonparametric Bayes model. For nonzero values of $\sigma_{\epsilon \nu}$, the classical and Bayesian LIV methods give approximate similar results. We note that the classical LIV model is correctly specified in all cases since the true number of groups is two. However, the performance of the nonparametric Bayesian LIV procedure is very encouraging.

As for the classical LIV model, in the nonparametric Bayesian LIV model we can test whether $\sigma_{\epsilon \nu} = 0$ (i.e. a test for endogeneity) easily by computing the fractions of the MCMC sample in which cases $\sigma_{\epsilon \nu} > 0$ and $\sigma_{\epsilon \nu} < 0$. We found for $\sigma_{\epsilon \nu} = 0$ that these fractions were close to 0.5, as they should be, and for both $\sigma_{\epsilon \nu} = 0.36$ and 0.79 that these were equal to 1 and 0, respectively. Hence,
a test for endogeneity can be computed straightforward as a byproduct of the MCMC output, using the above procedure based on the posterior \( P \)-values (see for a discussion Meng, 1994, or Sellke, Bayarri and Berger, 2001).

Table 7.2: Results main parameters for a skewed gamma distribution for the tree situations: \( \sigma_{\epsilon_v} = 0 \) (A), \( \sigma_{\epsilon_v} = 0.36 \) (B), and \( \sigma_{\epsilon_v} = 0.79 \) (C).

<table>
<thead>
<tr>
<th></th>
<th>( \beta_1 )</th>
<th>( \sigma^2_\epsilon )</th>
<th>( \sigma_{\epsilon_v} )</th>
<th>( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.00 (0.023)</td>
<td>1.01 (0.053)</td>
<td>-0.02 (0.087)</td>
<td>13 (3.05)</td>
</tr>
<tr>
<td>Bayes LIV</td>
<td>2.00 (0.049)</td>
<td>1.02 (0.053)</td>
<td>-0.02 (0.072)</td>
<td>2</td>
</tr>
<tr>
<td>LIV2</td>
<td>2.00 (0.043)</td>
<td>1.01 (0.053)</td>
<td>-0.02 (0.082)</td>
<td>3</td>
</tr>
<tr>
<td>LIV3</td>
<td>1.99 (0.031)</td>
<td>1.01 (0.088)</td>
<td>-0.02 (0.082)</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.14 (0.018)</td>
<td>0.95 (0.056)</td>
<td>0.37 (0.057)</td>
<td>11 (2.45)</td>
</tr>
<tr>
<td>Bayes LIV</td>
<td>2.00 (0.021)</td>
<td>1.00 (0.066)</td>
<td>0.37 (0.074)</td>
<td>2</td>
</tr>
<tr>
<td>LIV2</td>
<td>2.00 (0.025)</td>
<td>1.00 (0.066)</td>
<td>0.38 (0.071)</td>
<td>3</td>
</tr>
<tr>
<td>LIV3</td>
<td>1.99 (0.024)</td>
<td>1.00 (0.071)</td>
<td>0.38 (0.071)</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.32 (0.028)</td>
<td>0.74 (0.043)</td>
<td>0.81 (0.075)</td>
<td>14 (2.31)</td>
</tr>
<tr>
<td>Bayes LIV</td>
<td>1.99 (0.026)</td>
<td>1.01 (0.082)</td>
<td>0.81 (0.096)</td>
<td>2</td>
</tr>
<tr>
<td>LIV2</td>
<td>1.99 (0.031)</td>
<td>1.01 (0.088)</td>
<td>0.81 (0.096)</td>
<td>2</td>
</tr>
<tr>
<td>LIV3</td>
<td>1.99 (0.029)</td>
<td>1.00 (0.086)</td>
<td>0.80 (0.082)</td>
<td>3</td>
</tr>
</tbody>
</table>

In table 7.2 we present the results for a situation where the unobserved instrument has a continuous skewed gamma distribution with scale parameter 0.5 and shape 0.577 (i.e. its variance is 1.5). It can be seen that the proposed nonparametric Bayesian LIV model gives unbiased results in all cases. Its results are preferred to the classical LIV results for \( \sigma_{\epsilon_v} > 0 \). Although the classical LIV model with three categories is slightly more efficient than with two categories, it is still less efficient than the nonparametric Bayes LIV model. For the classical LIV model with \( k > 3 \) we found degenerate solutions in several runs, which indicates that \( k = 3 \) is more or less the “best” choice. It can be seen that the nonparametric Bayes model can adapt more easily to a situation where the true distribution of the instrument is not discrete but continuous. As before, when \( \sigma_{\epsilon_v} = 0 \) the OLS estimate is best and the standard LIV model
Chapter 7 A Nonparametric Bayesian LIV approach

gives more efficient results than the nonparametric Bayes model. Surprisingly, the estimated number of clusters \( k \) is for the skewed gamma distribution much lower for \( \sigma_{\epsilon\nu} = 0 \) and 0.36 than when the true distribution of \( \theta_{ij} \) is discrete with two categories in table 7.1. This can be expected if the two components of the discrete distribution are not far apart, in which case the sampled distribution of \( \theta \) resembles more a symmetric, unimodal distribution, with a flat top, which is approximated by a large number of support points drawn from a normal distribution. Apparently, a skewed gamma distribution is approximated by a mixture of normals with fewer support points. The proposed test to test for endogeneity was found to give satisfactory results (posterior \( P \)-values of 0.45, 1, and 1 for \( \sigma_{\epsilon\nu} = 0, 0.36, \) and 0.79).

Table 7.3: Results main parameters for a \( t \) distribution with six degrees of freedom for the tree situations: \( \sigma_{\epsilon\nu} = 0 \) (A), \( \sigma_{\epsilon\nu} = 0.36 \) (B), and \( \sigma_{\epsilon\nu} = 0.79 \) (C).

<table>
<thead>
<tr>
<th></th>
<th>( \beta_1 )</th>
<th>( \sigma^2_\epsilon )</th>
<th>( \sigma_{\epsilon\nu} )</th>
<th>( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.00 (0.006)</td>
<td>1.00 (0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes LIV</td>
<td>1.99 (0.027)</td>
<td>1.00 (0.016)</td>
<td>0.03 (0.066)</td>
<td>29 (42.7)</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.14 (0.005)</td>
<td>0.94 (0.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes LIV</td>
<td>1.99 (0.028)</td>
<td>1.00 (0.024)</td>
<td>0.38 (0.069)</td>
<td>15 (8.94)</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>2.32 (0.004)</td>
<td>0.74 (0.015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes LIV</td>
<td>2.00 (0.032)</td>
<td>1.00 (0.058)</td>
<td>0.80 (0.078)</td>
<td>19 (6.75)</td>
</tr>
</tbody>
</table>

In table 7.3 we present the results for the simulated \( t_6 \) distribution of the instruments. We found that a sample size of \( n = 1000 \) was not sufficient to estimate the model. In general, the MCMC chain did not converge, indicating non- or under-identification, which was immediately clear from convergence plots for (e.g.) \( \sigma_{\epsilon\nu} \). This can be expected because the \( t \) distribution is close to a normal distribution which is not identified, and a relatively large sample is needed to have a good representation of the tails of that distribution. The results presented in table 7.3 are for a total sample size of \( n = 10000 \) (obtained as, for instance, \( n = 10000 \) and \( m = 1 \), or \( n = 1000 \) and \( m = 10 \)).
before, we used 15 simulated datasets. We also estimated the synthetic data with the standard LIV model, but in all cases we found that the estimated Hessian matrix had several eigenvalues equal to zero, indicating that the model is not identified. This reveals that the standard LIV model is more sensitive to the distribution of the instruments when it is close to normal. From the results in Table 7.3 it becomes clear that the nonparametric Bayes LIV model yields approximate unbiased results, but the relatively large standard deviations indicate that the model is weakly identified. Examination of the iteration plots suggests that the chain has converged. Contrary to the results in the previous, the nonparametric Bayes estimates exhibit more variability for larger amounts of endogeneity, although part of this variability is expected to reduce when the number of simulated datasets is increased.

The simulation studies presented here illustrate that the nonparametric Bayes approach with a general distribution for the unobserved instrument is powerful in estimating linear models in presence of regressor-error correlations. We compared the new method with classical OLS and the LIV method proposed in the previous chapters and we examined two extreme cases, one case where the classical LIV method is correctly specified and one case which represents near identifiability. When the distribution of the instrument is truly discrete, the classical LIV approach performs best, but the nonparametric approach gives approximately similar results. When the true distribution of the instrument is continuous the nonparametric Bayes approach performs better, which illustrates its flexibility in adapting to the distribution of $\theta$. The standard LIV model could not be estimated when the unobserved instrument has a $t$ distribution. The nonparametric Bayesian LIV method, however, does give approximate unbiased results but the estimated standard deviations may be rather large. Besides, the results critically rely on the sample size used, which may present a problem for cross sectional applications. However, this may be less an issue in multilevel studies where typically several observations on a subject are available.

In the following we present the results for the multilevel model described in
section 7.3. Here we consider a situation with one endogenous regressor and a situation with two endogenous regressors. Given the problems found with the $t$ distribution for the simple model, we only present results for an unobserved instrument that has a discrete distribution and a continuous skewed distribution. This model cannot be estimated by the standard LIV model developed in the previous chapters.

### 7.4.2 Simulation results for the hierarchical model

We considered two situations for model (7.17): one in which case we have one endogenous regressor $z_1$, i.e. $l_1 = 1$, and a situation with two, $l_1 = 2$. In both cases we assumed the presence of one exogenous covariate $z_2$ ($l_2 = 1$), with a small effect ($\alpha$) on the endogenous covariates, one regressor $x$, and a constant, such that $\beta_i$ is of dimension two, for $i = 1, \ldots, n$. We took $n = 500$ and $m = 15$ in all cases. As before, we simulated 15 datasets and we considered three situations: no, moderate and severe endogeneity (the corresponding elements of $\Sigma_{\beta z_1}$ are 0, 0.36 and 0.79, respectively). We first present the results for one endogenous regressor, where the true distribution of $\theta_j$ is a (univariate) discrete distribution with two categories, subsequently we present the results for two endogenous covariates where the distribution of the unobserved instrument is a (bivariate) skewed gamma distribution. The estimates are compared with results from a standard hierarchical Bayes model as in Lenk et al. (1996).

#### Results for one endogenous regressor

The results are presented in table 7.4. We only present the results for the regression parameter $\beta_1$, the regression parameters $\gamma$ that correspond to the endogenous covariate $z_1$, and the estimated covariances $\Sigma_{\beta z_1}$. We found that for two of the 15 simulated datasets the estimated value for $k$ equaled 1 in all MCMC draws, in which case the model is not identified. These situations can be identified easily from examining iteration plots of $k$ or $\gamma$, see the figure in appendix 7D. The estimated values for the nonparametric Bayes model for $\gamma$ and $\Sigma_{\beta z_1}$ in table 7.4 are obtained after excluding the non-converged cases.
7.4 A simulation study

Table 7.4: Results main parameters for hierarchical model with one possible endogenous regressor tree situations: $\Sigma_{\beta_1} = 0 \times t_2$ (A), $\Sigma_{\beta_{z1}} = 0.36 \times t_2$ (B), and $\Sigma_{\beta_{z1}} = 0.79 \times t_2$ (C). The other true values are: $\beta_1 = 2$ and $\gamma_{11} = \gamma_{21} = 1$.

<table>
<thead>
<tr>
<th></th>
<th>A NPB LIV</th>
<th>Std HB</th>
<th>B NPB LIV</th>
<th>Std HB</th>
<th>C NPB LIV</th>
<th>Std HB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>2.01</td>
<td>(0.06)</td>
<td>1.99</td>
<td>(0.07)</td>
<td>2.02</td>
<td>(0.08)</td>
</tr>
<tr>
<td>$\gamma_{11}$</td>
<td>1.03</td>
<td>(0.06)</td>
<td>1.00</td>
<td>(0.03)</td>
<td>1.02</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\gamma_{21}$</td>
<td>0.99</td>
<td>(0.06)</td>
<td>1.00</td>
<td>(0.08)</td>
<td>1.29</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\Sigma_{\beta_{z1}}^{(11)}$</td>
<td>-0.05</td>
<td>(0.13)</td>
<td>0.34</td>
<td>(0.15)</td>
<td>0.70</td>
<td>(0.10)</td>
</tr>
<tr>
<td>$\Sigma_{\beta_{z1}}^{(21)}$</td>
<td>0.02</td>
<td>(0.13)</td>
<td>0.37</td>
<td>(0.20)</td>
<td>0.68</td>
<td>(0.08)</td>
</tr>
<tr>
<td>$k$</td>
<td>13</td>
<td>(3.70)</td>
<td>11</td>
<td>(2.74)</td>
<td>5</td>
<td>(1.30)</td>
</tr>
</tbody>
</table>

Two results are immediately clear. Firstly, the results for the main regression parameter $\beta_1$ are almost equal for the nonparametric Bayes model and the standard hierarchical Bayes model, regardless of whether a covariate $z$ is endogenous or not. Hence, the results for the main regression parameters $\beta$ are unaffected by the presence of an endogenous covariate $z$ (in this example, however, the covariates $z$ were generated independently from the regressors $x$, i.e. they are not collinear, which may potentially explain the result found). Secondly, as expected, when endogeneity is present, the estimated regression parameters $\gamma$ obtained from the standard hierarchical Bayes model are biased, as expected. It can be seen that the nonparametric Bayes model corrects for this and allows for unbiased estimation.

We note that the effective sample size used to estimate the distribution of the latent instrument in this simulation study is 500, whereas in the previous simulation studies the sample size was at least 1000. Furthermore, the regression
parameters $\beta_i$ are unobserved parameters and, hence, less information is available to estimate the distribution of the unobserved instrument. Hence, we typically observe larger standard deviations than for the results in table 7.1.

In all cases $\sigma^2_x$, the variance of the main regression equation for $y$, was estimated unbiasedly by both the nonparametric Bayes LIV model and the standard hierarchical Bayes model. Furthermore, although not given in table 7.4, the nonparametric Bayesian LIV model estimates for the heterogeneity variances of the regression parameters $\Sigma_{\beta_i}$ are approximately unbiased, regardless of the presence of endogenous covariates. Contrary, the heterogeneity variances are severely underestimated by the standard hierarchical Bayes model in presence of an endogenous covariate. This was also found in chapter 6 for the variance of the random intercept (for instance table 6.2).

It can be seen from the above results that the bias due to level-2 endogeneity in the standard hierarchical Bayes estimates for $\gamma$, and $\Sigma_{\beta_i}$, may be quite large. In the following subsection we present the results for a situation with two endogenous covariates.

**Results for two endogenous regressors**

Since there are a large number of parameters in this model, we only focus on the main parameters. As for the results in the previous subsection with one endogenous covariate, the results for the main regression equation are not affected by presence of endogeneity of some of the $z$’s (while assuming that the $z$’s and the $x$’s are not collinear), and both the nonparametric Bayes LIV model and the standard hierarchical Bayes model give approximate similar results. We therefore choose to report the results only for the elements of $\gamma$ ($\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}$) corresponding to the endogenous covariates and the the elements of $\Sigma_{\beta_{z_i}}$. Furthermore, we report the results for the estimated number of clusters $k$.

The results are presented in table 7.5. It can be seen that the nonparametric Bayes LIV model gives unbiased results for the regression parameters $\gamma$ and
7.4 A simulation study

Table 7.5: Results main parameters for hierarchical model with two possible endogenous regressors for tree situations: \( \Sigma_{\beta z_1} = 0 \times I_2 \) (A), \( \Sigma_{\beta z_1} = 0.36 \times I_2 \) (B), and \( \Sigma_{\beta z_1} = 0.79 \times I_2 \) (C). The other true values are: \( \gamma_{11} = \gamma_{21} = 1 \) and \( \gamma_{12} = \gamma_{22} = -1 \).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NPB LIV</td>
<td>Std HB</td>
<td>NPB LIV</td>
</tr>
<tr>
<td>( \gamma_{11} )</td>
<td>0.98 (0.05)</td>
<td>1.00 (0.03)</td>
<td>0.98 (0.05)</td>
</tr>
<tr>
<td>( \gamma_{12} )</td>
<td>-1.00 (0.05)</td>
<td>-1.00 (0.03)</td>
<td>-1.00 (0.05)</td>
</tr>
<tr>
<td>( \gamma_{21} )</td>
<td>0.99 (0.07)</td>
<td>0.99 (0.03)</td>
<td>0.99 (0.05)</td>
</tr>
<tr>
<td>( \gamma_{22} )</td>
<td>-0.99 (0.07)</td>
<td>-0.98 (0.03)</td>
<td>-1.02 (0.05)</td>
</tr>
<tr>
<td>( \Sigma_{\beta z_1}^{(11)} )</td>
<td>0.03 (0.13)</td>
<td>0.40 (0.12)</td>
<td>0.03 (0.09)</td>
</tr>
<tr>
<td>( \Sigma_{\beta z_1}^{(12)} )</td>
<td>0.03 (0.09)</td>
<td>0.37 (0.11)</td>
<td>0.03 (0.10)</td>
</tr>
<tr>
<td>( \Sigma_{\beta z_1}^{(21)} )</td>
<td>0.01 (0.17)</td>
<td>0.38 (0.08)</td>
<td>0.01 (0.08)</td>
</tr>
<tr>
<td>( \Sigma_{\beta z_1}^{(22)} )</td>
<td>0.01 (0.17)</td>
<td>0.40 (0.11)</td>
<td>0.01 (0.11)</td>
</tr>
<tr>
<td>( k )</td>
<td>23 (12.71)</td>
<td>18 (5.64)</td>
<td>21 (3.58)</td>
</tr>
</tbody>
</table>

the covariances between \((\eta_i, \xi_i)\) in (7.17) and (7.18), that induce the dependency between \(\beta_i\) and the elements of \(z_{11}\). Furthermore, the estimated values for \(\gamma\) using the standard hierarchical Bayes model are biased when the covariates \(z_i\) are not exogenous. For instance, for the case with severe endogeneity (C), this bias amounts to approximately 25%, which is quite severe.

We found that the estimates for the heterogeneity variance components\(^5\) \(\Sigma_{\beta \beta}\) are approximately unbiased for the nonparametric Bayes LIV model, but the

---

\(^5\)Not reported here.
standard hierarchical Bayes model underestimates the variances by about 40% when there is severe endogeneity.

7.5 Discussion nonparametric Bayesian LIV approach

In this chapter we presented preliminary findings on a very general approach to model endogeneity in single or multilevel models. We proposed a nonparametric Bayes approach to model the distribution of the latent instrument using a Dirichlet prior process. We considered two general multilevel models that may suffer from regressor-error dependencies. One advantage of using a Bayesian approach is that it can handle complex model structures in a straightforward manner through MCMC estimation. We illustrated that the nonparametric Bayes model for the unobserved instrument proposed in section 7.2 could be adapted to a more general setting in section 7.3 without too much effort. Although the technicalities surrounding Dirichlet prior processes may be demanding, it presents a flexible approach that can be extended and adapted easily to other situations with potential regressor-error dependencies.

The simulation results showed that the nonparametric Bayes LIV model gives unbiased results for a variety of settings. We compared the approach proposed in this chapter to the LIV model in chapter 3, and found that the nonparametric Bayes approach outperforms the classical LIV model for non-discrete distributions, because the Dirichlet process prior allows for full estimation of the distribution of the unobserved instrument. We saw that the model yields approximately unbiased results, even for the extreme case when the unobserved instrument has a $t$ distribution, although a large dataset needs to be available. Implementation of the nonparametric Bayes LIV model does not require a priori specification of the number of clusters, but this number is estimated as a by-product of the estimation. Similarly, we found for the random coefficients model in section 7.3 that the nonparametric Bayesian LIV approach can be successfully used to estimated the model parameters in presence of endogenous covariates. Importantly, the estimates for $\gamma$ obtained from the standard hierarchical Bayes model are strongly biased. In addition, the standard hierar-
chical Bayes model substantially underestimates the amount of heterogeneity in the regression coefficients. We showed that the nonparametric Bayesian LIV approach can handle situations with more than one endogenous regressor.

Although the results are promising, future research is needed to obtain more insight in using a Dirichlet prior process for the distribution of the unobserved instrument. The simulation studies presented here are informative, but limited to only three kinds of distributions for the unobserved instrument: a discrete distribution with two categories, a heavily skewed gamma distribution and a $t$ distribution with fat tails. We plan to investigate the properties of the nonparametric Bayes LIV model for a broader range of distributions. The results above suggest that when the distribution of the latent instrument is close to a normal distribution the results should be interpreted with caution, which was revealed by examining iteration plots of key parameters. We suggest to investigate convergence issues in detail in empirical applications, since simply relying on iterations plots may be too limiting (Cowles and Carlin, 1996, Brooks and Roberts, 1998). Besides, we plan to investigate whether the mixing of the MCMC chain can be improved and whether this depends on the amount of autocorrelation between subsequent MCMC draws. We found that the computational burden for the nonparametric Bayes LIV model is much larger than for the standard LIV model.

One interesting application for the nonparametric Bayes LIV approach is to combine the Dirichlet process prior with the Hausman-Taylor approach presented in subsection 6.3.3. The Hausman-Taylor approach can be applied to general random intercept models that suffer from level-two dependencies. Hausman and Taylor (1981) show that the multilevel structure of the data and prior knowledge on the exogeneity of part of the available regressors can be used to construct instrumental variables to estimate the regression parameters. This method has the advantage that no external instrumental variables are required. Furthermore, their approach allows for estimation of level-two (group-specific) variables as opposed to e.g. Mundlak’s approach or fixed-effects estimation. The Hausman-Taylor estimator, however, was shown to be limited
in its use when level-one regressor-error dependencies are present (see section 6.4). Importantly, in constructing the ‘internal’ instruments as proposed by Hausman and Taylor focus is only on whether or not certain regressors can be assumed independent of the random intercepts. Although this may be a valid assumption in many applications, their method does not address the strength of the obtained ‘internal’ instruments, and the method seems to be ad-hoc at this point. We illustrated this important aspect in section 6.6. Incorporation of a general distributed unobserved instrument in the Hausman-Taylor model can be done using the results in section 7.2, and may yield improved results when the proposed ‘internal’ instruments are weak or when possible level-one endogeneity is present.

Furthermore, our nonparametric Bayesian LIV approach can possibly handle situations where both level-1 and level-2 dependencies are present. Lousily spoken, the errors in (7.2) can be regarded as a result of two terms: (1) a level-2 specific error, and (2) a level-1 specific error. I.e., the ‘total’ errors are $\tilde{\epsilon}_{ij} = \alpha_i + \epsilon_{ij}$ and $\tilde{\nu}_{ij} = \tau_j + \nu_{ij}$. The variance-covariance matrix $\Sigma$ in (7.3) can be changed accordingly. Level-2 endogeneity arises when $E(\alpha_i \tau_j) \neq 0$, and level-1 endogeneity when $E(\epsilon_{ij} \nu_{ij}) \neq 0$. In both cases, $E(\tilde{\epsilon}_{ij} \tilde{\nu}_{ij}) \neq 0$, which is in form similar to the problem considered in section 7.2. It is interesting to investigate this extension, in particular given the conclusions in the previous chapter.

The standard LIV model in chapter 3 is identified when the unobserved instrument has at least two categories. When the unobserved instrument has one mean ($m = 1$), the parameters are not identified. In a Dirichilet process there is a positive probability that the number of different values for the latent instrument $\theta$ is less than the sample size $n_j$ (i.e. the $\theta$’s are clustered). Our simulation studies suggest that identification problems occur when the true distribution of the unobserved instrument gets close to normal. We found clear indications of lack of convergence of the MCMC output in such cases, suggesting that the nonparametric Bayes model ‘automatically’ points-out a non-identified solution, and the results should be discarded. Nevertheless, further research is required to investigate this conjecture. Observations made by Lewbel (1997)
or Carroll, Roeder and Wasserman (1999) may prove helpful. Related to this, we suspect that, if the error distribution $\nu_{ij}$ in (7.2) is non-normal, a normal distribution for the latent instrument may be possible.
Appendix 7A  The Dirichlet process

Ferguson (1973) introduces the Dirichlet process priors as a class of prior distributions for a set of probability distributions on a given sample space. The Dirichlet process is based on the Dirichlet distribution, which is given in definition 7A.1 (Ferguson, 1973).

Definition 7A.1 Let $Z_1, Z_2, \ldots, Z_k$ be independent random variables with $Z_i \sim \text{Gamma}(\alpha_i, 1)$, where $\alpha_i \geq 0$ for all $i$, and $\alpha_i > 0$ for some $i$. Let $Y_i = Z_i / \sum_j Z_j$.

Then distribution of $(Y_1, \ldots, Y_k)$ is a Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_k)$.

The probability density for a Dirichlet distribution is defined by

$$f(y_1, \ldots, y_k|\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^k y_i^{\alpha_i - 1},$$

(7A.1)

where $\alpha_1, \ldots, \alpha_k > 0$, $y_1, \ldots, y_k \geq 0$ and $\sum_i y_i = 1$. Let $\alpha = \sum_i \alpha_i$. The mean of a Dirichlet distribution is $E(Y_i) = \alpha_i / \alpha$ and the variance is $\text{var}(Y_i) = \alpha_i(\alpha - \alpha_i) / \alpha^2(\alpha + 1)$. The Dirichlet distribution is an extension of the Beta distribution ($k = 2$). The Dirichlet distribution can be used as a prior for the (discrete) probabilities (group sizes) in e.g. mixture models (see also property 3, Ferguson, 1973).

Ferguson’s Dirichlet process is extensively discussed in Ferguson (1973) and Antoniak (1974). Here we present definition 1 from Antoniak:

Definition 7A.2 Let $\Theta$ be a set and $\mathcal{A}$ be a $\sigma$-field of subsets of $\Theta$. Let $\nu$ be a finite, non-null, non-negative, finitely additive measure on $(\Theta, \mathcal{A})$. Now, a random probability measure $P$ on $(\Theta, \mathcal{A})$ is a Dirichlet process on $(\Theta, \mathcal{A})$ with parameter $\nu$, if for every $k = 1, 2, \ldots$ and measurable partition $B_1, \ldots, B_k$ of $\Theta$, the joint distribution of the random probabilities $(P(B_1), \ldots, P(B_k))$ is a Dirichlet distribution with parameters $(\nu(B_1), \ldots, \nu(B_k))$.

We write: $P \in \mathcal{D}(\nu)$. Ferguson obtains the following properties of the Dirichlet process:

1. If $P \in \mathcal{D}(\nu)$, and $A \in \mathcal{A}$, then $E(P(A)) = \nu(A) / \nu(\Theta)$;

2. If $P \in \mathcal{D}(\nu)$, and, conditional given $P$, $\theta_1, \theta_2, \ldots, \theta_n$ are an i.i.d. sample from $P$, then $P|\theta_1, \theta_2, \ldots, \theta_n \in \mathcal{D}(\nu + \sum_{i=1}^n \delta_{\theta_i})$, where $\delta_x$ is a measure giving mass one to point $x$;

3. If $P \in \mathcal{D}(\nu)$, then $P$ is almost surely discrete.

Furthermore, Ferguson (1973) shows that for a sample $X$ of size 1 from $P \in \mathcal{D}(\nu)$, $\mathbb{P}(X \in A) = \nu(A) / \nu(\Theta)$, for $A \in \mathcal{A}$.

In nonparametric Bayesian applications it is common to specify $\nu$ as $\nu = \alpha G_0$, where $G_0$ is a distribution, and $\alpha > 0$. The posterior distribution, which is conditional on
the data, that arises from a structure with a Dirichlet process prior, is known to be 
results on MCMC estimation, however, show that the full conditional distributions are 
fairly simple to use in parameter estimation.

**Appendix 7B**  
**Full conditionals: the simple multilevel model with general LIV**

**Full conditional for** $\Sigma$. For $\Sigma$ we take the inverted Wishart prior with parameters $\omega$ and $\Psi$, implying

$$p(\Sigma|\text{rest, data}) \propto (\det(\Sigma))^{-\tilde{n}/2} \exp \left( -\frac{1}{2} \sum_{i} \sum_{j} (z_{ij} - \mu_{z_{i},j}^{x})^{\prime} \Sigma^{-1} (z_{ij} - \mu_{z_{i},j}^{x}) \right) \times$$

$$\times |\Sigma|^{-(\omega+2+1)/2} \exp \left( -\frac{1}{2} \text{tr} (\Psi \Sigma^{-1}) \right) \times$$

$$\propto |\Sigma|^{-(\omega+\tilde{n}+2+1)/2} \exp \left( -\frac{1}{2} \text{tr} \left( \left( \sum_{i} \sum_{j} (z_{ij} - \mu_{z_{i},j}^{x})(z_{ij} - \mu_{z_{i},j}^{x})^{\prime} + \Psi \right) \Sigma^{-1} \right) \right) ,$$

i.e. the full conditional of $\Sigma$ is a inverted Wishart with parameters $\omega + \tilde{n}$ and 
$\left( \sum_{i} \sum_{j} (z_{ij} - \mu_{z_{i},j}^{x})(z_{ij} - \mu_{z_{i},j}^{x})^{\prime} + \Psi \right)$.  

**Full conditional for** $\beta$. The prior for $\beta$ is a bivariate normal distribution with mean 
$\mu_{\beta}$ and variance $\Sigma_{\beta}$. Because both likelihood and prior are normal distributions, 
we focuss on the ‘quadratic’ term (kernel) (for the sake of national convenience we 
suppress in the following the subscript $i, j$, and let $\tilde{\eta} = (x - \theta)$ and $\tilde{x} = (1, x)$)

$$(z - \mu_{z_{i}}^{x})^{\prime} \Sigma^{-1} (z - \mu_{z_{i}}^{x}) =$$

$$= \left( \begin{array}{c} y \\ x \end{array} \right) - \left( \begin{array}{c} \tilde{x} \beta \\ \theta \end{array} \right) \right)^{\prime} \left( \begin{array}{cc} \sigma^{(11)} & \sigma^{(12)} \\ \sigma^{(12)} & \sigma^{(22)} \end{array} \right) \left( \begin{array}{c} \left( y - \tilde{x} \beta \right) \\ \left( x - \tilde{x} \beta \right) \end{array} \right)$$

$$= \sigma^{(11)} (y - \tilde{x} \beta)^{\prime} (y - \tilde{x} \beta) + \sigma^{(12)} \tilde{\eta} (y - \tilde{x} \beta) + \sigma^{(12)} \tilde{\eta} (y - \tilde{x} \beta)^{\prime} + \tilde{\eta}^{2} \sigma^{(22)}$$

$$\times \sigma^{(11)} (-\beta^{\prime} \tilde{x}^{\prime} y - y \tilde{x} \beta + \beta^{\prime} \tilde{x}^{\prime} \tilde{x} \beta) - \sigma^{(12)} \tilde{\eta} \tilde{x} \beta - \sigma^{(12)} \tilde{\eta} \tilde{x} \beta - \sigma^{(12)} \tilde{\eta} \tilde{x} \beta + \tilde{n} \beta^{\prime} \tilde{x} \tilde{\eta},$$

$$= \sigma^{(11)} \beta^{\prime} \tilde{x}^{\prime} \tilde{x} \beta - \beta^{\prime} \tilde{x}^{\prime} \tilde{x} \beta (\sigma^{(11)} y + \sigma^{(12)} \tilde{\eta}) - (\sigma^{(11)} y + \sigma^{(12)} \tilde{\eta}) \tilde{x} \beta .$$

(7B.2)

By adding the subscripts and the kernel of the prior we get,
\[
\beta'|(\sigma^{(11)} \sum_{i,j} \tilde{z}_{ij} \tilde{x}_{ij} + \Sigma_{\beta}^{-1}) \beta - \beta'|(\sigma^{(11)} \sum_{i,j} \tilde{z}_{ij} (\sigma^{(11)} y_{ij} + \sigma^{(12)} \tilde{\eta}_{ij}) + \Sigma_{\beta}^{-1} \mu_{\beta}) + \\
-(\mu_{\beta} \Sigma_{\beta}^{-1} + \sum_{i,j} (\sigma^{(11)} y_{ij} + \sigma^{(12)} \tilde{\eta}_{ij}) \tilde{x}_{ij}) \beta.
\]

Let \( C = \sigma^{(11)} \sum_{i,j} \tilde{z}_{ij} \tilde{x}_{ij} + \Sigma_{\beta}^{-1} \). Now it follows that the full conditional for \( \beta \), given the other parameters and the data, is bivariate normal with mean

\[
C^{-1} \sum_{i,j} \tilde{z}_{ij} (\sigma^{(11)} y_{ij} + \sigma^{(12)} \tilde{\eta}_{ij}) + \Sigma_{\beta}^{-1} \mu_{\beta},
\]

and variance-covariance \( C^{-1} \). Note that if \( \Sigma_{\beta}^{-1} = 0 \) the mean of the full conditional for \( \beta \) is similar to an ordinary regression with a correction term for the endogeneity. If their is no endogeneity, i.e. \( \sigma^{(12)} = 0 \), the ‘standard’ regression model is obtained.

**Derivation of full conditional distribution of \( \theta_{ij} \).** In the following we derive the expressions for the components of (7.10).

**Derivation of \( h(\theta_{ij}|\text{rest, data}) \).** \( h(\theta_{ij}|\text{rest, data}) \propto p(z_{ij}|\theta_{ij}, \beta, \Sigma) g_{0}(\theta_{ij}|\mu_{x}, \sigma_{\theta}^{2}) \) can be computed in a similar way as the full conditional of \( \beta \). The kernel of the (structural) likelihood, where \( \tilde{\epsilon} = y - \tilde{x} \beta \), is

\[
(z - \mu_{\tilde{\epsilon}})' \Sigma^{-1}(z - \mu_{\tilde{\epsilon}}) = \\
= \sigma^{(11)} \tilde{\epsilon}' \tilde{\epsilon} + \sigma^{(12)} (x - \theta) \tilde{\epsilon} + \sigma^{(12)} (x - \theta) \tilde{\epsilon}' + (x - \theta)^2 \sigma^{(22)} \\
\propto -2 \sigma^{(12)} \theta \tilde{\epsilon} + \sigma^{(22)} \theta^2 - 2 \sigma^{(22)} x \theta.
\]

Adding the subscript \( i,j \) and the kernel of \( g_{0} \) yields

\[
\left( \sigma^{(22)} + \frac{1}{\sigma_{g}^{2}} \right) \tilde{\theta}_{ij}^2 - 2 \left( \sigma^{(22)} x_{ij} + \sigma^{(12)} \tilde{\epsilon}_{ij} + \frac{\mu_{x}}{\sigma_{g}} \right) \tilde{\theta}_{ij},
\]

from which it follows that \( h(\theta_{ij}|\text{rest, data}) \) is a normal density with mean

\[
\tilde{C}^{-1} \left( \sigma^{(22)} x_{ij} + \sigma^{(12)} \tilde{\epsilon}_{ij} + \frac{\mu_{x}}{\sigma_{g}} \right),
\]

and variance \( \tilde{C}^{-1} \), with \( \tilde{C} = \sigma^{(22)} + \frac{1}{\sigma_{g}^{2}} \). Note: when \( \sigma_{g}^{2} \to \infty \), the location of the distribution \( h(\theta_{ij}) \) is estimated by \( x_{ij} \) and an ‘endogeneity’ correction \( (\sigma^{(12)}/\sigma^{(22)}) \tilde{\epsilon}_{ij} \) (since \( \sigma^{(12)} = -\sigma_{\epsilon\epsilon}/|\Sigma| \) and \( \sigma^{(22)} = \sigma_{\epsilon\epsilon}^{2}/|\Sigma| \)).
Appendix 7B  Full conditionals: the simple multilevel model with general LIV

Derivation of \( q_0 \). \( q_0 \) is proportional to \( \alpha \int p(z_{ij} | \theta_{ij}, \beta, \Sigma) dG_0(\theta_{ij} | \mu_\theta, \sigma_\theta^2) \). In the following we present the steps necessary to integrate \( \theta_{ij} \) out of the likelihood function. We focus on the quadratic term of the reduced form likelihood (7.6). This derivation is conditional on all parameters (except \( \theta_{ij} \)) and the data. We drop \( ij \)-subscript for notational convenience and use \( \tilde{z} = (v - \tilde{\mu}_0, x)' \) and \( \tilde{\mu}_\tilde{z} = \theta v \), where \( v = (\beta_1, 1)' \).

Then,

\[
(\tilde{z} - \tilde{\mu}_\tilde{z})' \Omega^{-1} (\tilde{z} - \tilde{\mu}_\tilde{z}) + \frac{(\theta - \mu_\theta)^2}{\sigma_\theta^2} = \tilde{z}' \Omega^{-1} \tilde{z} - \theta v' \Omega^{-1} v + \frac{\theta^2}{\sigma_\theta^2} - 2 \frac{\mu_\theta}{\sigma_\theta^2} + \frac{\mu_\theta^2}{\sigma_\theta^2}
\]

By letting \( \kappa_1 = v' \Omega^{-1} \tilde{z} \), \( \kappa_2 = v' \Omega^{-1} v + \frac{1}{\sigma_\theta^2} \), and rearranging terms (the terms not involving \( \theta \) or \( \tilde{z} \) are dropped in the factor of proportionality), we get

\[
\kappa_2 \theta^2 - 2 \left( \frac{\mu_\theta}{\sigma_\theta^2} + \kappa_1 \right) \theta + \tilde{z}' \Omega^{-1} \tilde{z} \tag{7B.5}
\]

which is equal to (take \( \kappa_3 = \kappa_1 + \frac{\mu_\theta}{\sigma_\theta^2} \))

\[
\kappa_2 \left( \theta - \frac{\kappa_3}{\kappa_2} \right)^2 - \frac{\kappa_3^2}{\kappa_2} + \tilde{z}' \Omega^{-1} \tilde{z} \tag{7B.6}
\]

The first term integrates to one. The remaining terms involve matrices and vectors and we have to be careful with taking squares, transposing and taking inverses. Let \( A = \kappa_2^{-1} \), now (7B.6) is equal to

\[
\tilde{z}' \Omega^{-1} \tilde{z} - \kappa_3 A \kappa_3 = \tilde{z}' \Omega^{-1} \tilde{z} - \left( \tilde{z}' \Omega^{-1} v + \frac{\mu_\theta}{\sigma_\theta^2} \right) A \left( v' \Omega^{-1} \tilde{z} + \frac{\mu_\theta}{\sigma_\theta^2} \right) \propto -\tilde{z}' \Omega^{-1} v A \frac{\mu_\theta}{\sigma_\theta^2} - \frac{\mu_\theta}{\sigma_\theta^2} A v' \Omega^{-1} \tilde{z} - \tilde{z}' \Omega^{-1} v A v' \Omega^{-1} \tilde{z} + \left( \Omega^{-1} - \Omega^{-1} v A v' \Omega^{-1} \right) \tilde{z} - \tilde{z}' \Omega^{-1} v A \frac{\mu_\theta}{\sigma_\theta^2} - \frac{\mu_\theta}{\sigma_\theta^2} A v' \Omega^{-1} \tilde{z}. \tag{7B.7}
\]

Let \( B = \Omega^{-1} - \Omega^{-1} v A v' \Omega^{-1} = \left[ \Omega + v \sigma^2 v' \right]^{-1} \), where we used expression (2-66b) from Greene (2000). The expression in (7B.7) is the kernel of a multivariate normal density\(^6\) with variance-covariance matrix \( B^{-1} = \Omega + v \sigma^2 v' \) and mean \( B^{-1} \Omega^{-1} v A \frac{\mu_\theta}{\sigma_\theta^2} \).

\(^6\)The general form of a multivariate normal distribution with mean \( V^{-1} \mu \) and variance-covariance \( V^{-1} \) is: \( (x - V^{-1} \mu)' V (x - V^{-1} \mu) = x' V x - \mu' V \mu = \mu' V^{-1} \mu. \)
This latter expression can be simplified to \( \nu \mu_g \). This is not immediately clear, but substituting \( A = \sigma^2_g - \sigma^2_g \nu' \left[ \Omega + \nu \sigma^2_g \nu' \right]^{-1} \nu \sigma^2_g \) and rearranging terms gives the desired result. Hence, \( q_0 \) is proportional to \( \alpha N_{ij} \), where \( N_{ij} \) is the density \(^7\) of a bivariate normal distribution at \( \tilde{z}_{ij} \) with mean \( \nu \nu_g \) and variance-covariance \( \Omega + \nu \sigma^2_g \nu' \), which is equal to

\[
B \Sigma B' + \nu^2 \begin{bmatrix} \beta_1 & \beta_0 \\ \beta_0 & 1 \end{bmatrix} = \\
\begin{bmatrix} \beta_1^2 \sigma^2_g + \sigma_x^2 + 2 \beta_1 \sigma_{ev} + \sigma_e^2 & \beta_1 (\sigma^2_g + \sigma_x^2) + \sigma_{ev} \\
\beta_1 (\sigma^2_g + \sigma_x^2) + \sigma_{ev} & \sigma_e^2 + \sigma_g^2 \end{bmatrix}.
\]

(7B.8)

**Derivation of \( q_j \).** This quantity is proportional to \( n_j^\top \) times \( p(z_{ij} | \tilde{\theta}, \beta, \Sigma) \), \( j = 1, ..., n^\top \), where \( n^\top \) is the number of different \( \tilde{\theta} \)'s when \( \theta_{ij} \) is removed.

**Remixing \( \tilde{\theta} \).** Using the expression for the posterior distribution of \( \tilde{\theta} \), it follows that the derivation of the remixing distribution for \( \tilde{\theta} \) is more or less similar to the derivation of \( h(\theta_{ij}) \) above. We have

\[
\sum_{i, j \in J_l} \left( \tilde{\epsilon} - \bar{x} \right)^\top \begin{bmatrix} \sigma^{(11)} & \sigma^{(12)} \\ \sigma^{(12)} & \sigma^{(22)} \end{bmatrix} \left( \tilde{\epsilon} - \tilde{\theta} \right) \propto \\
\sum_{i, j \in J_l} \left\{-2 \sigma^{(12)} \tilde{\epsilon} + \sigma^{(22)} \tilde{\epsilon}^2 - 2 \sigma^{(22)} x \tilde{\theta} \right\},
\]

(7B.9)

and by adding the kernel of \( g_0 \) and the subscripts \( ij \) we obtain

\[
\left( \tilde{n}_l \sigma^{(22)} + \frac{1}{\sigma^2} \right) \tilde{\theta}^2_l - 2 \left( \sigma^{(22)} \sum_{i, j \in J_l} x_{ij} + \sigma^{(12)} \sum_{i, j \in J_l} \tilde{z}_{ij} + \frac{\mu_g}{\sigma^2} \right) \tilde{\theta}_l.
\]

(7B.10)

It follows that the (full conditional) posterior remixing distribution for \( \tilde{\theta}_l \) is a normal distribution with mean \( C^{-1}(\sigma^{(22)} \sum_{i, j \in J_l} x_{ij} + \sigma^{(12)} \sum_{i, j \in J_l} \tilde{z}_{ij} + \frac{\mu_g}{\sigma^2}) \) and variance \( C^{-1} \), where \( C = \tilde{n}_l \sigma^{(22)} + \frac{1}{\sigma^2} \) and \( J_l \) is defined in subsection 7.2.2. When \( \sigma^2 \to \infty \) and there is no endogeneity (\( \sigma^{(12)} = 0 \), then the mean of the full conditional distribution for \( \tilde{\theta}_l \) is equal to the sample mean of \( x_{ij} \) of the observations in cluster \( l \). The

\(^7\)This result looks very much like the results of a more simple case: \( \int \int f(y | \mu, \sigma^2) g(\mu | \mu_0, \sigma^2_0) d\mu \), where \( y \) has a normal distribution with mean \( \mu \) and variance \( \sigma^2 \) and \( \mu \) has a normal distribution with mean \( \mu_0 \) and variance \( \sigma^2_0 \), of which we know that this integral yields \( y | \sigma^2, \mu_0, \sigma^2_0 \sim N(\mu_0, \sigma^2 + \sigma^2_0) \).
Appendix 7C  Full conditionals: the hierarchical model with general LIV

variance is in this case equal to $\sigma^2 / \bar{n}_i$.

**Full conditional for $\mu_g$.** From (7.15) we obtain

$$
\frac{\sum_l (\tilde{\theta}_l - \mu_g)^2}{\sigma^2_g} + \frac{(\mu_g - \mu_0)^2}{\sigma^2_0} \propto \mu^2_g \left( \frac{1}{\sigma^2_0} + \frac{\bar{n}}{\sigma^2_g} \right) - 2 \left( \frac{\sum_l \tilde{\theta}_l}{\sigma^2_g} + \frac{\mu_0}{\sigma^2_0} \right) \mu_g,
$$

from which it follows that the full conditional for $\mu_g$ is a normal distribution with mean $C^{-1} (\sum_l \tilde{\theta}_l / \sigma^2_g + \mu_0 / \sigma^2_0)$ and variance $C^{-1}$, with $C = (1/\sigma^2_0 + \bar{n} / \sigma^2_g)$. Note: if $\sigma^2_0 \to \infty$ than the mean of the full conditional distribution is computed as $\sum_l \tilde{\theta}_l / \bar{n}$ and its variance as $\sigma^2_g / \bar{n}$.

**Full conditional for $\sigma^2_g$.** The derivation is similar to the derivation of the full conditional distribution for $\mu_g$. We use (7.16), hence

$$
\frac{1}{\left(\sqrt{\sigma^2_g}\right)^2} \exp \left( -\frac{1}{2} \sum_{j=1}^{\bar{n}} \frac{(\tilde{\theta}_j - \mu_g)^2}{\sigma^2_g} \right) \frac{1}{(\sigma^2_g)^{c+1}} \exp \left( -\frac{d}{\sigma^2_g} \right) =
$$

$$
\frac{1}{(\sigma^2_g)^{(c+\bar{n}/2+1)}} \exp \left( -\frac{1}{2} \sum_{j=1}^{\bar{n}} (\tilde{\theta}_j - \mu_g)^2 + d \right),
$$

from which it follows that the full conditional distribution for $\sigma^2_g$ is an inverted gamma distribution with parameters $c + \bar{n}/2$ and $1/2 \sum_{j=1}^{\bar{n}} (\tilde{\theta}_j - \mu_g)^2 + d$.

**Full conditional for $\alpha$.** The full conditional distribution for the dispersion parameter $\alpha$ of the Dirichlet process can be obtained using the results in e.g. West (1992) or Escobar and West (1995).

Appendix 7C  Full conditionals: the hierarchical model with general LIV

**Full conditional for $\sigma^2$ and $\beta_i$.** Assuming conjugate prior densities, the full conditional distributions for $\sigma^2$ and $\beta_i$ can be derived easily using standard results from combining a normal likelihood with a inverted gamma distribution, and with a multivariate normal distribution with mean (7.20) and variance-covariance (7.21), respectively.
Full conditional distribution for $\Lambda$. The prior distribution for $\Lambda$ is an uninformative $(k + l_i)$ dimensional inverted Wishart distribution with parameters $c$ and $D$. Let $h_i = (\beta_i', z_i')'$ and $\mu_{h_i} = ((\gamma_i + \gamma z_i)', (\theta_i + \alpha z_i)')'$. Now

$$p(\Lambda|\text{rest, data}) \propto \left\{ \prod_{i=1}^{n} \det(\Lambda) \frac{1}{2} \exp \left[ -\frac{1}{2} \left( (h_i - \mu_{h_i})' \Lambda^{-1} (h_i - \mu_{h_i}) \right) \right] \right\} \times$$

$$\times \det(\Lambda)^{-\frac{c+k+l_i+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \left( D \Lambda^{-1} \right) \right\},$$

(7C.1)

from which it follows that the full conditional for $\Lambda$ is also an inverted Wishart$(k+l_i)$-distribution with parameters $n + c$ and

$$\sum_{i=1}^{n} (h_i - \mu_{h_i})(h_i - \mu_{h_i})' + D.$$

Full conditional for the matrix $\gamma_i$. We use the vectorized system in (7.22) and (7.23). The quadratic term, under normality of the errors $(\tilde{\eta}_i', \tilde{\xi}_i')'$, of this system is given by (we drop the subscript $i$ for the moment)

$$\begin{pmatrix} \tilde{\beta} - \tilde{z}_0\tilde{\gamma}_0 \end{pmatrix}' \begin{pmatrix} \Lambda^{(11)} & \Lambda^{(12)} \\ \Lambda^{(21)} & \Lambda^{(22)} \end{pmatrix} \begin{pmatrix} \tilde{\beta} - \tilde{z}_0\tilde{\gamma}_0 \\ \tilde{z}_10 \end{pmatrix}$$

$$\propto$$

(7C.2)

$$\tilde{\gamma}_0\Lambda^{(11)}\tilde{z}_0\tilde{\gamma}_0 - \tilde{\gamma}_0\Lambda^{(12)}\tilde{z}_{10} + \Lambda^{(11)}\tilde{\beta} - (\tilde{\beta}'\Lambda^{(11)} + \tilde{z}_{10}'\Lambda^{(21)})\tilde{z}_0\tilde{\gamma}_0.$$

The prior for $\tilde{\gamma}_0$ is a $(l_1+l_2+1)$-variate normal distribution with mean $m_\gamma$ and variance $V_\gamma$. Adding the subscript $i$ and the prior kernel we obtain

$$\tilde{\gamma}_0\left( \sum_i z_{0i}'\Lambda^{(11)}z_{0i} \right) \tilde{\gamma}_0 - \tilde{\gamma}_0' \left( \sum_i z_{0i}' \left( \Lambda^{(12)}z_{10i} + \Lambda^{(11)}\tilde{\beta}_i \right) \right) +$$

$$- \left( \sum_i \left( \tilde{\beta}_i'\Lambda^{(11)} + z_{10i}'\Lambda^{(21)} \right) z_{0i} \right) \tilde{\gamma}_0 + \tilde{\gamma}_0' V_\gamma^{-1} \tilde{\gamma}_0 - m_\gamma' V_\gamma^{-1} m_\gamma$$

$$=\tilde{\gamma}_0' \left[ \sum_i z_{0i}'\Lambda^{(11)}z_{0i} + V_\gamma^{-1} \right] \tilde{\gamma}_0 - \tilde{\gamma}_0' \left[ \sum_i z_{0i}' \left( \Lambda^{(12)}z_{10i} + \Lambda^{(11)}\tilde{\beta}_i \right) + V_\gamma^{-1} m_\gamma \right] +$$

$$- \left[ \sum_i \left( \tilde{\beta}_i'\Lambda^{(11)} + z_{10i}'\Lambda^{(21)} \right) z_{0i} + V_\gamma^{-1} m_\gamma \right] \tilde{\gamma}_0,$$

(7C.3)

from which it follows that the full conditional distribution for $\tilde{\gamma}_0$ is a $(l_1+l_2+1)$-variate normal distribution with mean
\[ \hat{C}^{-1} \left[ \sum_i \hat{z}_{i0} ' \left( \Lambda^{(12)} \hat{z}_{10i} + \Lambda^{(11)} \hat{\rho}_i \right) + V_y^{-1} \gamma \right], \quad (7C.4) \]

and variance-covariance \( \hat{C}^{-1} \), where \( \hat{C} = \sum_i \hat{z}_{i0} ' \Lambda^{(11)} \hat{z}_{0i} + V_y^{-1} \).

**Full conditional for the matrix** \( \alpha \). Similarly, the quadratic term of the multivariate normal distribution of the system in (7.26) is given by

\[ \left( \hat{\beta}_0 \right)' \left[ \Lambda^{(11)} \hat{\beta}_0 + \Lambda^{(12)} \right] \left( \hat{\beta}_0 \right) \]

\[ \propto \hat{\alpha} ' z_2 (A^{(22)} z_2 + \alpha) - \hat{\alpha} ' z_2 \left( \Lambda^{(21)} \hat{\beta}_0 + \Lambda^{(22)} \right) \hat{z}_{10} \hat{z}_{10}. \]

\[ (7C.5) \]

Adding the kernel of the prior for \( \hat{\alpha} \) (a \( l_1 l_2 \)-variate normal distribution) and the subscripts \( i \), we find that (similar as for \( \hat{\lambda} \)) the full conditional density function for \( \hat{\alpha} \) is a multivariate normal distribution with mean (let \( \hat{C} = \sum_i \hat{z}_{2i} ' \Lambda^{(22)} \hat{z}_{2i} + V_y^{-1} \)),

\[ \hat{C}^{-1} \left[ \sum_i \hat{z}_{2i} ' \left( \Lambda^{(21)} \hat{\beta}_0 + \Lambda^{(22)} \hat{z}_{10i} \right) + V_y^{-1} \gamma \right], \quad (7C.6) \]

and variance \( \hat{C}^{-1} \).

**Full conditional distribution for** \( \theta_i \). In the following we derive the probabilities \( q_{0(i)}, q_{1(i)}, \ldots, q_{\tilde{n}-(i)} \) and the density \( h(\theta_i | \text{data, rest}) \).

**Derivation of** \( q_0 \). We proceed in the same way as in the derivation of \( q_0 \) for the nonparametric Bayes LIV model in section 7.2. The reduced form is equal to (we drop subscript \( i \) for the moment)

\[ \beta = \gamma c + \gamma_1 \theta + (\gamma_1 \alpha + \gamma_2) z_2 + \gamma_1 \xi + \eta, \]

\[ z_1 = \theta + \alpha z_2 + \xi. \]

\[ (7C.7) \]

Define \( u = \gamma_1 \xi + \eta \), which is equal to

\[ u = \begin{pmatrix} \gamma_1 \xi + \eta \\ \xi \end{pmatrix} = B \begin{bmatrix} \eta \\ \xi \end{bmatrix} \]

\[ (7C.8) \]

where
Define $\tilde{A} = \text{var}(\mu', \xi') = B A B'$. The likelihoods for structural and reduced form are equal, since the Jacobian of transformation is 1. The following is conditional on all the other parameters. Let $\beta_0 = \beta - \gamma_c - (\gamma_1 \alpha + \gamma_2)z_2$ and $z_{10} = z_1 - \alpha z_2$, and define $h = (\beta_0, z_{10})$, a $(k + l_1) \times 1$ vector. Furthermore, let $\nu = (\nu_1', \nu_2')$, a $(k + l_1) \times l_1$-matrix, and $\mu_k = \nu \theta$. $q_0 \propto A(h) = \rho \int p(h|\theta, \gamma_c, \alpha, \Lambda, \text{data}) dG_0(\theta|\mu_\theta, V_\theta)$, i.e. $q_0$ is proportional to the density that results when $\theta$ is integrated out. We examine the kernel of the product of these two multivariate normal densities:

\[
(h - \mu_k)' \tilde{A}^{-1} (h - \mu_k) + (\theta - \mu_\theta)' V_\theta^{-1} (\theta - \mu_\theta) \propto h' \tilde{A}^{-1} h - h' \tilde{A}^{-1} \nu \theta - (\nu \theta)' \tilde{A}^{-1} h + (\nu \theta)' \tilde{A}^{-1} \nu \theta + \nu' V_\theta^{-1} \nu + \theta' V_\theta^{-1} \theta + \mu_\theta' V_\theta^{-1} \mu_\theta.
\]

Define

\[
\kappa_1 = \nu' \tilde{A}^{-1} h
\]
\[
\kappa_2 = \nu' \tilde{A}^{-1} \nu + V_\theta^{-1}
\]
\[
\kappa_3 = \kappa_1 + V_\theta^{-1} \mu_\theta.
\]

Then it can be shown that (7C.9) is equal to

\[
(\theta - \kappa_2^{-1} \kappa_3)' \nu \tilde{A}^{-1} (\theta - \kappa_2^{-1} \kappa_3) - \kappa_2^{-1} \kappa_3 h' \tilde{A}^{-1} h.
\]

The first term is the kernel of a multivariate normal distribution and integrates to 1, and we only need to focus on the latter two terms. In letting $\tilde{A} = \kappa_2^{-1}$ and $\tilde{B} = \tilde{A}^{-1} - \tilde{A}^{-1} \nu \tilde{A}^{-1} \nu$, we find that

\[
h' \tilde{A}^{-1} h - \kappa_2^{-1} \kappa_3 = h' \tilde{B} h - h' \tilde{A}^{-1} \nu \tilde{A}^{-1} \mu_\theta - \mu_\theta' V_\theta^{-1} \mu_\theta = \nu \mu_\theta
\]

from which it can be seen that $A(h)$ is proportional to a $(k + l_1)$-variate normal distribution with variance $\tilde{B}^{-1}$ and mean $\tilde{B}^{-1} \tilde{A}^{-1} \nu \tilde{A}^{-1} \mu_\theta = \nu \mu_\theta$ (this last equality is not immediately clear but follows after completing all products).

*Derivation of $q_j$.* $q_j$ is proportional to $\tilde{n}_j$ times the density $p(h_j|\tilde{\theta}_j, \gamma_c, \alpha, \Lambda, \text{data})$, $j = 1, ..., \tilde{n}$, which is a $(k + l_1)$-dimensional multivariate normal density with mean $((\gamma_c + \gamma_1 \tilde{\theta}_j + (\gamma_1 \alpha + \gamma_2)z_2)', (\tilde{\theta}_j + \alpha z_2)')'$ and variance $\tilde{A}$.

*Derivation of $h(\theta_i|\text{rest, data})$.* $h(\theta_i|\text{rest, data})$ is proportional to
Appendix 7C  Full conditionals: the hierarchical model with general LIV

\[ f(h_i|\theta_i, \alpha, \gamma_0, \Lambda, \text{data})g_0(\theta_i|\mu_\theta, V_\theta), \]

where \( g_0 \) is the probability density function of a multivariate normal distribution. We first consider the kernel of \( f \) (drop the subscript \( i \), let \( \beta_0 = \beta - \gamma_e - \gamma_z \), and \( z_{10} = z_1 - \alpha z_2 \)). We have

\[
\left( \begin{array}{c} \beta_0 \\ z_{10} - \theta \end{array} \right) \sim \left( \begin{array}{c} \Lambda^{(11)} \\ \Lambda^{(12)} \end{array} \right) \left( \begin{array}{c} \beta_0 \\ z_{10} - \theta \end{array} \right) \propto \theta' \Lambda^{(22)} \theta - \theta' (\Lambda^{(21)} \beta_0 + \Lambda^{(22)} z_{10}) - (\beta_0 \Lambda^{(12)} + z_{10}' \Lambda^{(22)}) \theta, \tag{7C.13} \]

and adding the kernel of the distribution \( g_0 \) gives

\[
\theta' \Lambda^{(22)} \theta - \theta' (\Lambda^{(21)} \beta_0 + \Lambda^{(22)} z_{10}) - (\beta_0 \Lambda^{(12)} + z_{10}' \Lambda^{(22)}) \theta + \theta' V_\theta^{-1} \theta - \mu_\theta' V_\theta^{-1} \theta - \theta' V_\theta^{-1} \theta' V_\theta^{-1} \theta' V_\theta^{-1} \theta', \tag{7C.14} \]

\[
\theta' (\Lambda^{(22)} + V_\theta^{-1}) \theta - \theta' (\Lambda^{(21)} \beta_0 + \Lambda^{(22)} z_{10} + V_\theta^{-1}) + (\beta_0 \Lambda^{(12)} + z_{10}' \Lambda^{(22)} + \mu_\theta' V_\theta^{-1} \theta), \tag{7C.15} \]

from which it can be seen that (write \( \tilde{\theta} = \Lambda^{(22)} + V_\theta^{-1} \)) \( h(\theta_i|\text{rest, data}) \) is a normal density with mean

\[
\tilde{\theta}^{-1} \left( \Lambda^{(21)} \beta_0 + \Lambda^{(22)} z_{10\theta} + V_\theta^{-1} \mu_\theta \right) \]

and variance-covariance \( \tilde{\theta}^{-1} \).

**Full conditional for \( \mu_\theta \).** We assume a multivariate normal prior density for \( \mu_\theta \) with mean \( m_\mu \) and variance \( V_\theta \). Using the normal kernel of the prior and \( G_0 \), and the cluster structure of the Dirichlet process, we have

\[
\sum_{l=1}^{\bar{n}} (\tilde{\theta}_l - \mu_\theta)' V_\theta^{-1} (\tilde{\theta}_l - \mu_\theta) + (\mu_\theta - m_\mu)' V_\mu^{-1} (\mu_\theta - m_\mu) \propto \mu_\theta' \left( \bar{n} V_\theta^{-1} + V_\mu^{-1} \right) \mu_\theta - \mu_\theta' \left( V_\theta^{-1} \sum_{l=1}^{\bar{n}} \tilde{\theta}_l + V_\mu^{-1} m_\mu \right) + \mu_\theta' \left( \tilde{\theta}_l' V_\theta^{-1} + m_\mu' V_\mu^{-1} \right) \mu_\theta, \]

\[
\mu_\theta' \left( \bar{n} V_\theta^{-1} + V_\mu^{-1} \right) \mu_\theta - \mu_\theta' \left( V_\theta^{-1} \sum_{l=1}^{\bar{n}} \tilde{\theta}_l + V_\mu^{-1} m_\mu \right) + \mu_\theta' \left( \tilde{\theta}_l' V_\theta^{-1} + m_\mu' V_\mu^{-1} \right) \mu_\theta.
\]
from which it can be seen that the full conditional distribution for \( \mu_\theta \) is a multivariate normal distribution with mean \( \tilde{C}^{-1} \left( V^{-1}_\theta \sum l=1^\tilde{n} \tilde{\theta}_l + V^{-1}_\mu m_\mu \right) \) and variance-covariance matrix \( \tilde{C}^{-1} \), where \( \tilde{C} = \tilde{n} V^{-1}_\theta + V^{-1}_\mu \).

**Full conditional distribution for** \( V_\theta \). Assuming an inverted Wishart density function with parameters \( (\tau V, \Upsilon V) \) as prior, and using similar arguments as for \( \mu_\theta \), we obtain the following full conditional distribution

\[
p(V_\theta|\text{data, rest}) \propto \left( \prod_{l=1}^{\tilde{n}} |V_\theta|^{-\frac{\tilde{n}}{2}} \exp \left\{ -\frac{1}{2} \left( \tilde{\theta}_l - \mu_\theta \right)' V^{-1}_\theta \left( \tilde{\theta}_l - \mu_\theta \right) \right\} \times \right. \\
\times |V_\theta|^{-\frac{\tau V + \tilde{n}}{2}} \exp \left[ -\frac{1}{2} \text{tr} \left( \Upsilon V V^{-1}_\theta \right) \right] \\
\left. \propto |V_\theta|^{-\frac{\tau V + \tilde{n} + l_1}{2}} \exp \left[ -\frac{1}{2} \text{tr} \left\{ \sum_{l=1}^{\tilde{n}} \left( \tilde{\theta}_l - \mu_\theta \right) \left( \tilde{\theta}_l - \mu_\theta \right)' + \Upsilon V \right\} V^{-1}_\theta \right], \right(7C.16)
\]

from which it follows that the full conditional for \( V_\theta \) is an inverted Wishart distribution of dimension \( l_1 \) with parameters \( \tau V + \tilde{n} \) and \( \sum_{l=1}^{\tilde{n}} (\tilde{\theta}_l - \mu_\theta)(\tilde{\theta}_l - \mu_\theta)' + \Upsilon V \).

**Full conditional for** \( \rho \). As before, we refer to West (1992) or Escobar and West (1995) to construct the full conditional distribution for the ‘dispersion’ parameter \( \rho \).
Appendix 7D Iteration plots

Figure 7D.1: Iteration plots $\gamma_{11}$ and $k$ for dataset no. 1 that failed to converge (with table 7.4).