Kotlarski with a Factor Loading

Arthur Lewbel* Boston College

April 2020

Abstract

This note extends the Kotlarski (1967) Lemma to show exactly what is identified when we allow for an unknown factor loading on the common unobserved factor. Potential applications include measurement error models and panel data factor models.

Suppose we observe, or can identify from data, the joint distribution of two random variables Y and X. Suppose that

$$\begin{aligned} Y &= cV + W \\ X &= V + U \end{aligned}$$

where U, V, and W are unobserved, mutually independent real valued random variables, and c is a nonzero constant. If we knew that $c = 1$, we could apply Kotlarski's (1967) Lemma to show that the distributions of U, V, and W are point identified. Since its introduction into the econometrics literature by Li and Vuong (1998), Kotlarski's Lemma has been widely applied. Prominent examples include Bonhomme and Robin (2010) and Cunha, Heckman, and Schennach (2010).

This note extends Kotlarski's Lemma to the case of unknown c. Most of this extension is a one line proof, making use of a theorem by Reiersøl (1950), and much of the rest exploits bounds like those of Frisch (1934). Given the age of these references, it's a bit surprising that Theorem 1 below hadn't been previously made explicit.

This extension is useful because many economic models have error structures of the above type, where V is an unobserved factor that affects outcomes Y and X, and c is a factor loading. Here are two examples of such models:

Example 1: Measurement Error in Linear Regression: Suppose X is the observed, mismeasured version of some unobserved true variable X^*, so $X = X^* + U$ where U is measurement

*JEL codes: C14, C30. Keywords: Kotlarski, deconvolution, factor models, measurement error. Corresponding Author: Arthur Lewbel, Department of Economics - Maloney 315, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://sites.google.com/bc.edu/arthur-lewbel/

error. Consider the regression model $Y = a + bX^* + e$. The goal would be to identify the coefficients a and b and the distributions of the true X^*, the measurement error U, and the model error e. Letting $V = X^*$, $W = a + e$, $a = E(W)$, and $b = c$, we then get that this model is equivalent to our framework of $X = V + U$ and $Y = cV + W$. Theorem 1 below identifies c, and the distributions of U, V, and W, which then identifies all the features of the model, noting that $a = E(W)$ and $e = W - E(W)$.

Example 2: Fixed T Panel Data Models: Consider the panel data model $Y_{it} = g_t(Z_{it}) + \lambda_t V_i + \varepsilon_{it}$, with unknown functions g_t, unobserved individual specific effects V_i, unknown constant factor loadings λ_t, and idiosyncratic errors ε_{it}. For the moment, assume $g_t(Z_{it})$ is just a constant α_t. The goal is then to identify α_t, the distribution of the individual specific effects V_i, each factor loading λ_t, and the distributions of the idiosyncratic effects ε_{it} in each period t. Impose the free normalization $\lambda_1 = 1$, and consider any other time period s. If we define $U_i = \alpha_1 + \varepsilon_{i1}$, $W_i = \alpha_s + \varepsilon_{is}$, and $\lambda_s = c$, we get that $\alpha_t + \lambda_t V_i + \varepsilon_{it}$ equals $V + U$ in period $t = 1$ and equals $cV + W$ in period $t = s$. We may then apply Theorem 1 below, followed by $\alpha_1 = E(U_i)$, $\varepsilon_{i1} = U_i - E(U_i)$, $\alpha_s = E(W_i)$, and $\varepsilon_{is} = W_i - E(W_i)$. This then gives α_t, λ_t, and the distributions of V_i and of ε_{it} for every observed time period t. Finally, if instead of constants α_t we had functions $g_t(Z_{it})$ for observed covariates Z_{it}, we could repeat the analysis conditioning on Z_{i1} and Z_{is}, and replace the constructions of α_{i1} and α_{is} with conditional means, conditioning on Z_{i1} and Z_{is}, respectively. Evdokimov (2010) uses the Kotlarski Lemma to identify a similar panel structure without factor loadings.

A number of special cases of this paper's result already appear in the literature, referring either directly to the system $Y = cV + W$ and $X = V + U$, or to the equivalent measurent error model of example 1 above. For example, the factor loading c (or equivalently, the mismeasured regressor coefficient b) is known to be point identified when one of the components is asymmetrically distributed. This fact is used by estimators proposed in Lewbel (1997) and Erickson and Whited (2002). Other papers assume more generally that the factor loading is point identified by higher moments, but do not explicitly characterize when that is possible (or what is identifiable in that case). Examples include Bonhomme and Robin (2010), Fruehwirth, Navarro, and Takahashi (2016), and Navarro and Zhou (2017). Generalizations of Kotlarski's lemma to models with more components (including some restricted factor loadings) include Székely and Rao (2000) and Li and Zheng (2019).

ASSUMPTION A1: We observe the joint distribution of two real valued, nondegenerate random variables Y and X.

ASSUMPTION A2: $X = V + U$ and $Y = cV + W$, where U, V, and W are mutually independent, real valued, random variables with unknown distributions. The unknown constant c is real valued, finite, and nonzero.

ASSUMPTION A3: Either the median or the mean of V is zero. The characteristic functions of U, V, and W do not vanish.

Assumption A3 can be relaxed. Kotlarski's Lemma requires some location normalization, and imposing it on the mean or median of V will be convenient. Evdokimov and White (2012)

provide alternative conditions under which the characteristic functions of U, V, and/or W can have zeros.

Kotlarski's Lemma assumes $c = 1$. We assume $c \neq 0$ because, if $c = 0$ then trivially we can only identify the distributions of W and of $V + U$. Moreover, we can immediately tell if $c = 0$, because in that case the distributions of X and Y will be independent.

For any random variable R, let $\mu_r = E(R)$ if this mean exists, let $\sigma_r^2 = var(R)$ if this variance exists, and let $\sigma_{xy} = cov(X, Y)$ if this covariance exists.

THEOREM 1: Let Assumptions A1, A2, and A3 hold. Then either

a. The constant c and the distributions of U, V, and W are all point identified, or

b. Y and X are jointly normally distributed. If this latter case holds, then c is interval identified by $c \in \left(\sigma_{xy}/\sigma_x^2, \sigma_y^2/\sigma_{xy}\right)$, and for every value of c in this interval, there exists a corresponding unique normal distribution for U, V, and for W.

Theorem 1 shows that the Kotlarski Lemma extends to point identification with an unknown factor loading c, as long as Y and X are not jointly normal. Otherwise, in the case of normality, the sharp identified sets for c, U, V, and W are obtained from a simple, easily calculated interval. In practice, given observations of Y and X, one could test for joint normality, to see which of the two cases applies.

This note concludes with a proof of Theorem 1. As pointed out earlier, most of the proof simply applies results by Reiersøl (1950) and Frisch (1934).

PROOF of Theorem 1: In the regression measurement error model of example 1 above, Theorem 1 of Reiersøl (1950) shows that b is point identified if and only if Y and X are not jointly normal. This is also a special case of Theorem 1 of Schennach and Hu (2013). Given the equivalence of example 1 to Assumption A2, this means that c is identified if and only if Y and X are not jointly normal.

In the case where Y and X are not jointly normal, so c is point identified, we have $Y/c = V + W/c$. We can then apply Kotlarski's Lemma to the observed joint distribution of Y/c and X to identify the distributions of U, V, and W/c, and hence also identify the distribution of W.

In the case where Y and X are jointly normal, U, V, and W must each be normal, meaning that their distributions are identified up to unknown means and variances, which are finite. Given the location normalization $\mu_v = 0$, the means of U and W are identified by $\mu_u = \mu_x$ and $\mu_w = \mu_y$. The remaining step now borrows heavily from the Frisch (1934) bounds on mismeasured linear regression. From observed second moments of the data we have $\sigma_y^2 = c^2\sigma_v^2 + \sigma_w^2$, $\sigma_x^2 = \sigma_v^2 + \sigma_u^2$, and $\sigma_{xy} = c\sigma_v^2$, which provides three equations in the four unknown constants σ_u^2, σ_w^2, σ_v^2, and c. The only constraints on the parameter values of normal distributions are that σ_u^2, σ_w^2, and σ_v^2 must all be positive. The equation $\sigma_{xy} = c\sigma_v^2$ means that the sign of c equals the sign of σ_{xy} to ensure $\sigma_v^2 > 0$. Then $\sigma_u^2 > 0$ requires $\sigma_x^2 - \sigma_{xy}/c > 0$ and $\sigma_w^2 > 0$ requires $\sigma_y^2 - c\sigma_{xy} > 0$. Therefore, either $\sigma_{xy} > 0$ and $\sigma_{xy}/\sigma_x^2 < c < \sigma_y^2/\sigma_{xy}$, or $\sigma_{xy} < 0$ and $\sigma_y^2/\sigma_{xy} < c < \sigma_{xy}/\sigma_x^2$. Either way c lies in the interval between σ_{xy}/σ_x^2 and σ_y^2/σ_{xy}. Finally, given any c that lies in this interval, there's a corresponding

unique distribution for U, V, and W that satisfy the assumptions, given by $V \sim N\left(0, \sigma_{xy}/c\right)$, $U \sim N\left(\mu_x, \sigma_x^2 - \sigma_{xy}/c\right)$, and $W \sim N\left(\mu_y, \sigma_y^2 - c\sigma_{xy}\right)$.

Bonhomme, S. and J. - M. Robin (2010), "Generalized Non-Parametric Deconvolution with an Application to Earnings Dynamics," The Review of Economic Studies, 77, 491–533.

Fruehwirth, J. C., S. Navarro, and Y. Takahashi (2016), "How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects," Journal of Labor Economics 34:4, 979-1021

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitiveand Noncognitive Skill Formation," Econometrica, 78, 883–931.

Erickson, T. and T. M. Whited, (2002), "Two-step GMM estimation of the errors-in-variables model using high-order moments," Econometric Theory, 18(3), 776-799.

Evdokimov, K. (2010), "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," Unpublished Manuscript.

Evdokimov, K. and H. White (2012), "Some Extensions of a Lemma of Kotlarski," Econometric Theory, 28(4), 925-932.

Frisch, R. (1934), "Statistical confluence analysis by means of complete regression systems," Vol. 5, Universitetets Økonomiske Instituut.

Lewbel, A. (1997), "Constructing Instruments for Regressions With Measurement Error When No Additional Data are Available, With an Application to Patents and R&D," Econometrica, 65(5), 1201-1213.

Li, S. and X. Zheng, (2019), "A Generalization of Lemma 1 in Kotlarski (1967)," Rice University unpublished manuscript.

Kotlarski, I. I. (1967), "On characterizing the gamma and normal distribution," Pacific Journal of Mathematics, 20, 69–76.

Li, T. and Q. Vuong (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," Journal of Multivariate Analysis, 65, 139–165.

Navarro, S. and J. Zhou, (2017) "Identifying agent's information sets: An application to a lifecycle model of schooling, consumption and labor supply," Review of Economic Dynamics, 25, 58-92.

Reiersøl, O. (1950), "Identifiability of a linear relation between variables which are subject to error," Econometrica, 18, 375-389.

Schennach, S. M. and Y. Hu (2013), "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information," Journal of the American Statistical Association, 108, 177-186.

GJ Székely, G. J. and C. R. Rao (2000), "Identifiability of distributions of independent random variables by linear combinations and moments" Sankhyā: The Indian Journal of Statistics, Series A, 62(2), 193-202.