

Penalized Least Squares Methods for Latent Variables Models*

A discussion of the papers

by Susanne Schennach and by Victor Chernozhukov

Stéphane Bonhomme

CEMFI, Madrid

December 2010

Abstract

In this note, we propose a least squares method with ℓ^1 penalty (based on the “Lasso”) to estimate models with latent variables. Our approach addresses the high dimensionality of these models, due to the presence of unknown distribution functions. It builds on a recent proposal by Bunea, Tsybakov, Wegkamp and Barbu (2010, *Annals of Statistics*) that uses penalized least squares for density estimation. We apply the method to a simple measurement error model. The extension to more general latent variables models raises a number of issues that we briefly discuss.

JEL codes: C13, C14.

Keywords: Latent variables models, Lasso, penalization.

*This note is a discussion of papers prepared for the World Congress of the Econometric Society, August 2010. I thank Gary Chamberlain and Chris Hansen for stimulating discussions. All errors are my own.

1 Introduction

The papers by Susanne Schennach and Victor Chernozhukov that appear in this volume synthesize some important contributions that these authors have made in different areas of econometrics: latent variables models for the former, and ℓ^1 -penalized estimation methods for the latter. This note shows that these two areas of research are related, and proposes a penalization approach to estimate models with latent variables.

Latent variables models. Part of the recent work of Susanne Schennach aims at providing nonparametric identification results in Latent Variables Models (LVM hereafter). As an important example, Hu and Schennach (2008) provide conditions under which all latent distributions are nonparametrically identified in nonlinear latent variables models that satisfy conditional independence restrictions. These results represent significant improvements in a literature that has so far mostly focused on linear models (Kotlarski, 1967, Székely and Rao, 2000).

Schennach's paper focuses on measurement error models, where the true regressor is an unobserved latent variable. However, the techniques she discusses may also be used in models with a group structure and panel data models (where the group fixed effects are the latent variables), or to dynamic decision problems (with unobserved states). For example, using similar techniques, Hu and Shum (2010) provide conditions under which structural dynamic models are nonparametrically identified under Markovian assumptions on the dynamics of unobserved state variables.

Although the recent literature has made important progress on nonparametric point-identification of LVM under economically plausible assumptions, the estimation side is far less developed. In additive models, nonparametric estimators have been proposed that rely on the use of characteristic functions (Horowitz and Markatou, 1996, Bonhomme and Robin, 2010). Recently, Evdokimov (2010) has used a similar idea in a nonlinear panel data regression model with fixed effects.

When models are nonlinear, however, simple characteristic-function based estimation is often impossible. A general strategy consists in modelling the unknown distributions parametrically, while letting the dimension of the parameter space increase with the sample size. The properties of *sieve* estimation approaches are now well understood in many contexts (Chen, 2007). Sieve maximum likelihood has been advocated in the context of measurement error models (Hu and Schennach, 2008) or nonlinear panel data models (Bester and Hansen, 2007), for example. One difficulty with the method of sieves, however, is the nonlinearity of the estimation problem which, combined with the high dimensionality of latent variables models, raises non-trivial computational issues.

ℓ^1 penalized least squares (Lasso). The paper by Victor Chernozhukov reviews a seemingly very different area of econometrics. The focus is on linear regression models with many regressors. In this context, least squares methods behave poorly, and some form of penalization of the estimates is called for.

In a seminal contribution, Tibshirani (1996) proposed to add an ℓ^1 penalization term to the least squares criterion. When the regression of interest is $y_i = \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i$, the *Lasso*

estimate is defined as follows:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N \left(y_i - \sum_{j=1}^J \beta_j x_{ij} \right)^2 + \lambda_N \sum_{j=1}^J |\beta_j|, \quad (1)$$

where $\lambda_N \geq 0$ is a penalty parameter, and where $\beta = \{\beta_1, \dots, \beta_J\}$.

The Lasso has become a major area of statistical research. There are two main reasons for this popularity. First, the Lasso estimate $\widehat{\beta}$ is the solution of a convex programming problem that can be computed efficiently.¹ Second, the solution typically exhibits *sparsity*, meaning that many elements $\widehat{\beta}_j$ are exactly zero. This is due to the presence of the absolute value in (1), and contrasts with other penalization schemes. For example, Ridge regression, which adds a quadratic penalty to the least squares criterion, typically yields many coefficients close to zero but not exactly equal to zero. Sparsity of the Lasso solution is attractive, as it delivers parsimonious, interpretable results. Together with Alexandre Belloni, Chernozhukov has contributed to the rapidly expanding Lasso literature, by introducing post-Lasso estimation (Belloni and Chernozhukov, 2010a), and extending the Lasso idea to the quantile regression framework (Belloni and Chernozhukov, 2010b).

The Lasso is a useful approach in economic applications also. In the paper that appears in this volume, Chernozhukov uses a macroeconomic regression as main illustration. He takes as an example economic growth, for which many (potential) determinants have been considered, and proposes a robust test of the convergence hypothesis. Other areas of application that he mentions are nonlinear regression (where the unknown regression function is approximated in a large dictionary of functions), and the many instruments problem (Belloni *et al.*, 2010).

Combining ideas from the these lines of research. This note proposes an ℓ^1 -penalized least squares approach to estimate models with latent variables. The need for penalization arises because of the high dimensionality of LVM, due to the presence of unknown distribution functions (i.e., infinite-dimensional parameters). Similarly to Lasso in a linear regression context, the particular estimator we consider is computationally convenient and delivers parsimonious results.

The proposed approach builds on a recent paper by Bunea *et al.* (2010), where an ℓ^1 -penalized least squares approach was introduced for density estimation. We apply the method to a simple measurement error model, and provide a small numerical illustration on simulated data. Extending the idea to more general LVM raises various issues that we briefly discuss in the conclusion.

2 A simple latent variable model

We start by discussing various strategies to estimate latent variables models. We will focus the discussion on a simple measurement error model.

Measurement error. Suppose that we are interested in documenting the relationship between an outcome Y and a covariate X^* . Suppose, however, that we do not observe X^* , but only an imperfect measure X . We make two assumptions.

¹For example, using the Least Angle Regression algorithm of Efron *et al.* (2004) and Friedman *et al.* (2010), which delivers the full path of Lasso estimates for λ_N taking any non-negative value.

Assumption 1 Y is independent of X given X^* .

Assumption 1 is quite common in nonlinear measurement error models. It amounts to assuming that the error-contaminated regressor X does not contain information about Y , other than that contained in the true regressor X^* .

Assumption 2 The distribution function of X given X^* is known.

The rest of the analysis will remain unchanged if instead of Assumption 2 we assume that a consistent estimate of the distribution function of X given X^* is available. This could be because the researcher disposes of an auxiliary dataset with observations on X and X^* (Chen, Hong and Tamer, 2005). Alternatively, this could be because independent repeated measures of X^* are available. We will come back to the repeated measures example in the conclusion.

Let f_Z ($f_{Z|W}$) be a generic notation for the distribution function of Z (or Z given W). The following identity, which is a direct implication of Assumption 1, will be useful:

$$\begin{aligned} f_{Y,X}(y, x) &= \int f_{Y|X, X^*}(y|x, x^*) f_{X^*}(x^*) f_{X|X^*}(x|x^*) dx^* \\ &= \int f_{Y|X^*}(y|x^*) f_{X^*}(x^*) f_{X|X^*}(x|x^*) dx^* \\ &= \int f_{Y, X^*}(y, x^*) f_{X|X^*}(x|x^*) dx^*. \end{aligned} \quad (2)$$

Note that, in this expression, $f_{X|X^*}$ is known by Assumption 2, while f_{Y, X^*} is unknown.

Penalized likelihood. Let $\{y_i, x_i\}$, $i = 1, \dots, N$, denote an i.i.d. sample from (Y, X) . A popular approach is to postulate a (possibly high-dimensional) parametric model for f_{Y, X^*} , depending on some parameter vector $\gamma \in \Gamma$. The (penalized) maximum likelihood estimate of γ is then given by:

$$\hat{\gamma} = \operatorname{argmax}_{\gamma \in \Gamma} \sum_{i=1}^N \log \left(\int f_{Y, X^*}(y_i, x^*; \gamma) f_{X|X^*}(x_i|x^*) dx^* \right). \quad (3)$$

Various choices for the parameter space Γ lead to several commonly used estimation strategies:

- If Γ is taken to be independent of the sample size, $\hat{\gamma}$ is a standard parametric maximum likelihood (ML) estimate.
- If the number of parameters in Γ_N increases with N , $\hat{\gamma}$ is a sieve ML estimator (Shen, 1997).
- The choice $\Gamma_N = \left\{ \gamma \in \mathbb{R}^J, \sum_{j=1}^J \gamma_j^2 \leq t_N \right\}$, where t_N is a function of the sample and γ_j denotes the j th element of γ , corresponds to ML with quadratic penalty (Ridge).
- Lastly, if $\Gamma_N = \left\{ \gamma \in \mathbb{R}^J, \sum_{j=1}^J |\gamma_j| \leq t_N \right\}$, then $\hat{\gamma}$ is an ML estimate with ℓ^1 penalty.

An attractive feature of penalized likelihood approaches is their ability to deal with a large model space, i.e. a flexible specification for f_{Y,X^*} . For example, as in standard Lasso, the ℓ^1 penalty tends to select parsimonious, interpretable specifications out of the large class of models available. This flexibility is appealing in the context of LVM, where the “true” model is often unclear *ex-ante*.

On the negative side, computation of penalized likelihood estimators is often not straightforward. To see this, consider the Lagrangian that corresponds to the ℓ^1 -penalized estimator:

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\operatorname{argmin}} - \sum_{i=1}^N \log \left(\int f_{Y,X^*}(y_i, x^*; \gamma) f_{X|X^*}(x_i|x^*) dx^* \right) + \lambda_N \sum_{j=1}^J |\gamma_j|, \quad (4)$$

where λ_N is a Lagrange multiplier which depends on the sample size.² The minimization in (4) does not take the form of a simple programming problem in general. Efficient algorithms have been proposed in some special cases. See for example Hastie and Park (2007) for generalized linear models. However, the computation problem is much harder than for standard Lasso, especially when the model space is large. The next section proposes a method for estimating LVM that preserves the computational simplicity of Lasso.

3 Penalized Least Squares density estimation

3.1 The estimator

Let $\{\Psi_k(y, x^*), k = 1, \dots, K\}$ denote a (large) dictionary of functions, possibly non-orthogonal.³ Then, let us specify:

$$f_{Y,X^*}(y, x^*) = \sum_{k=1}^K a_k \Psi_k(y, x^*),$$

where $\{a_k\}$ are scalar parameters to be estimated.

We have, using (2):

$$\begin{aligned} f_{Y,X}(y, x) &= \int f_{Y,X^*}(y, x^*) f_{X|X^*}(x|x^*) dx^* \\ &= \sum_{k=1}^K a_k \int \Psi_k(y, x^*) f_{X|X^*}(x|x^*) dx^* \\ &= \sum_{k=1}^K a_k \varphi_k(y, x), \end{aligned} \quad (5)$$

where $\varphi_k(y, x) = \int \Psi_k(y, x^*) f_{X|X^*}(x|x^*) dx^*$ are known functions.

Equation (5) shows that the problem of recovering the parameters of the latent distribution f_{Y,X^*} from the distribution function of the data is linear. Here our aim is to devise an estimation method that exploits the model’s linearity.

²Note that (4) penalizes all elements γ_j in the same way, irrespective of the behavior of $\partial f_{Y,X^*}(y_i, x^*; \gamma) / \partial \gamma_j$. Our estimator in the next section will allow the penalization to depend on j .

³For example, $\{\Psi_k\}$ might contain bivariate normal densities with different means and covariance matrices, as well as orthogonal polynomials (Chebyshev, Hermite...) up to a very high order.

In what follows, we will work with square-integrable functions, and denote as $\|g\|_2^2 = \iint g(y, x)^2 dydx$ the squared ℓ^2 norm. Note that $\{a_k\}$ minimizes the ℓ^2 distance between the distribution function of the data $f_{Y,X}$ and its parametric approximation:⁴

$$\{a_k\} = \operatorname{argmin}_{\{a_k\}} \left\| f_{Y,X} - \sum_{k=1}^K a_k \varphi_k \right\|_2^2. \quad (6)$$

As the ℓ^2 norm of $f_{Y,X}$ does not depend on $\{a_k\}$, (6) is equivalent to:⁵

$$\begin{aligned} \{a_k\} &= \operatorname{argmin}_{\{a_k\}} -2\mathbb{E} \left(\sum_{k=1}^K a_k \varphi_k(Y, X) \right) + \left\| \sum_{k=1}^K a_k \varphi_k \right\|_2^2 \\ &= \operatorname{argmin}_{\{a_k\}} -2\mathbb{E} \left(\sum_{k=1}^K a_k \varphi_k(Y, X) \right) + \sum_{k=1}^K \sum_{\ell=1}^K a_k a_\ell \iint \varphi_k(y, x) \varphi_\ell(y, x) dydx. \end{aligned} \quad (7)$$

Given a random sample and taking empirical counterparts in (7) (as in Birgé and Massart, 1997), a least squares estimate of $\{a_k\}$ is thus:

$$\{\hat{a}_k\} = \operatorname{argmin}_{\{a_k\}} -\frac{2}{N} \sum_{i=1}^N \sum_{k=1}^K a_k \varphi_k(y_i, x_i) + \sum_{k=1}^K \sum_{\ell=1}^K a_k a_\ell \iint \varphi_k(y, x) \varphi_\ell(y, x) dydx. \quad (8)$$

ℓ^1 -penalization. When $\{\Psi_k\}$ is a high-dimensional, non-orthogonal, dictionary, the solution of (8) may be very imprecise. In this context, penalizing the solution may, by driving the coefficient estimates toward zero, improve finite-sample performance dramatically.

Here we follow Bunea *et al.* (2010) and add an ℓ^1 penalty to (8). The penalized estimates $\{\hat{a}_k\}$ are given by:

$$\begin{aligned} \{\hat{a}_k\} &= \operatorname{argmin}_{\{a_k\}} -\frac{2}{N} \sum_{i=1}^N \sum_{k=1}^K a_k \varphi_k(y_i, x_i) + \sum_{k=1}^K \sum_{\ell=1}^K a_k a_\ell \iint \varphi_k(y, x) \varphi_\ell(y, x) dydx \\ &\quad + \lambda_N \sum_{k=1}^K \|\varphi_k\|_\infty |a_k|, \end{aligned} \quad (9)$$

where the presence of the weights:

$$\|\varphi_k\|_\infty = \sup_{(y,x)} \varphi_k(y, x)$$

ensures that the estimator is invariant to the scale of Ψ_k , and where $\lambda_N \geq 0$ depends on the sample size.

⁴With some abuse of notation, $\{a_k\}$ on the left-hand side of (6) denotes the true value of the parameter.

⁵To see this, note that the ℓ^2 scalar product of $\sum_{k=1}^K a_k \varphi_k$ and $f_{Y,X}$ is:

$$\iint \sum_{k=1}^K a_k \varphi_k(y, x) f_{Y,X}(y, x) dydx = \mathbb{E} \left(\sum_{k=1}^K a_k \varphi_k(Y, X) \right).$$

Bunea *et al.* (2010) refer to:

$$\hat{f}_{Y,X} = \sum_{k=1}^K \hat{a}_k \varphi_k$$

as the SParse Density ESTimator (SPADES) of $f_{Y,X}$. The present discussion shows that the same idea can be used to estimate the joint density of the outcome and the true regressor in our measurement error model, as:

$$\hat{f}_{Y,X^*} = \sum_{k=1}^K \hat{a}_k \Psi_k.$$

Bunea *et al.* (2010) provide conditions under which SPADES $\hat{f}_{Y,X}$ yields minimax adaptive density estimates. We conjecture that \hat{f}_{Y,X^*} should satisfy similar optimality properties, although proving this conjecture exceeds the scope of this note.

Remark 1. The optimality results on ℓ^1 -penalized density estimation do not require that the true model (i.e., $\{a_k\}$) be sparse, but rather that there exist a sparse model that is close enough (in ℓ^2 norm) to the true model. This suggests that the proposed method should perform well when the distribution function f_{Y,X^*} can be well approximated by a few elements of the dictionary $\{\Psi_k\}$.

Remark 2. When the family $\{\varphi_k\}$ is orthogonal, SPADES has an analytical solution (it coincides with “soft thresholding”). For density estimation, which is the setting that Bunea and coauthors considered, the researcher may be willing to use an orthogonal basis of functions. However, in the simple measurement error model that we consider in this note, finding a dictionary $\{\Psi_k\}$ such that $\{\varphi_k\}$ be orthogonal is not a simple problem.⁶ The ability of SPADES to deal with non-orthogonal families of functions is thus an essential feature of the approach for the types of applications we have in mind.

Remark 3. A nice feature of (9) is that SPADES can be computed using a standard Lasso routine. For this, one needs to compute a matrix C of integrals with elements:

$$C_{k\ell} = \iint \varphi_k(y, x) \varphi_\ell(y, x) dy dx, \quad (k, \ell) \in \{1, \dots, K\}^2. \quad (10)$$

Moreover, stacking all a_k 's into one vector a , and denoting as φ the $K \times 1$ vector with elements $\frac{1}{N} \sum_{i=1}^N \varphi_k(y_i, x_i)$, (9) can be written as:

$$\hat{a} = \underset{a}{\operatorname{argmin}} -2\varphi' a + a' C a + \lambda_N \sum_{k=1}^K \|\varphi_k\|_\infty |a_k|, \quad (11)$$

or, equivalently:

$$\hat{a} = \underset{a}{\operatorname{argmin}} (P' a - P^\dagger \varphi)' (P' a - P^\dagger \varphi) + \lambda_N \sum_{k=1}^K \|\varphi_k\|_\infty |a_k|, \quad (12)$$

⁶This is because $\varphi_k(y, x) = \int \Psi_k(y, x^*) f_{X|X^*}(x|x^*) dx^*$ is a (non-additive) convolution of Ψ_k and $f_{X|X^*}$.

where P is any full-column rank matrix such that $C = PP'$, and P^\dagger denotes a generalized inverse of C .⁷ Algorithms that efficiently compute Lasso estimates in (12) are now publicly available.⁸

Remark 4. One limitation of SPADES is that the solution is not necessarily a valid density. In practice, when $\{\Psi_k\}$ is a set of densities, it may be useful to enforce the constraints: $a_k \geq 0$ and $\sum_{k=1}^K a_k = 1$. Constructing a modified Lasso algorithm that takes into account these constraints in the quadratic programming problem seems of interest.

Parameters of interest. Once an estimate \widehat{f}_{Y,X^*} is available, one can recover interesting features of the relationship between Y and X^* . In the illustration below, we will focus on the conditional expectation $\mathbb{E}(Y|X^*)$, which can be estimated at any point x^* as:

$$\begin{aligned} \widehat{\mathbb{E}}(Y|X^* = x^*) &= \frac{\int y \widehat{f}_{Y,X^*}(y, x^*) dy}{\int \widehat{f}_{Y,X^*}(y, x^*) dy} \\ &= \frac{\sum_{k=1}^K \widehat{a}_k \int y \Psi_k(y, x^*) dy}{\sum_{k=1}^K \widehat{a}_k \int \Psi_k(y, x^*) dy}, \end{aligned} \quad (13)$$

where $\{\widehat{a}_k\}$ is given by (9).

Relatedly, suppose we have a model for the conditional expectation:

$$\mathbb{E}(Y|X^* = x^*) = m(x^*; \theta),$$

where θ is a finite-dimensional parameter. A natural estimate of θ is given by:

$$\begin{aligned} \widehat{\theta} &= \operatorname{argmin}_{\theta} \iint (y - m(x^*; \theta))^2 \widehat{f}_{Y,X^*}(y, x^*) dy dx^* \\ &= \operatorname{argmin}_{\theta} \sum_{k=1}^K \widehat{a}_k \iint (y - m(x^*; \theta))^2 \Psi_k(y, x^*) dy dx^*. \end{aligned} \quad (14)$$

We will compute the estimates (13) and (14) in the illustration below.

3.2 Numerical illustration

As an illustration, we consider the following simple measurement error model:

$$\begin{cases} Y &= \frac{1}{\theta} \log(1 + \exp(\theta X^*)) + V \\ X &= X^* + U, \end{cases}$$

where V and U are independent given X^* (Assumption 1), and U follows a standard normal distribution (known to the researcher by Assumption 2). In the data generating process, we take $\theta = 1/2$, V and X^* standard normal, and we take U , V and X^* independent.

⁷The equivalence between (11) and (12) follows from the fact that $(I_K - PP^\dagger)\varphi = 0$ (where I_K denotes the $K \times K$ identity matrix), because φ belongs to the range of P .

⁸See <http://www-stat.stanford.edu/~tibs/lasso.html> for a fast Matlab implementation of the algorithm of Friedman *et al.* (2010), which we use in the illustration.

Let ϕ denote the standard normal density. We use the following dictionary of functions:

$$\Psi_{k,\ell}(y, x^*) = \frac{1}{\sigma_k} \phi\left(\frac{y - \mu_k}{\sigma_k}\right) \times \frac{1}{\sigma_\ell} \phi\left(\frac{x^* - \mu_\ell}{\sigma_\ell}\right), \quad (k, \ell) \in \{1, \dots, 25\}^2,$$

where $\mu_k \sim \mathcal{N}(0, 1)$, and $\sigma_k \sim \frac{1}{10} + \chi_1^2$. We approximate the integrals in (10) by importance sampling (1000 draws).

Figure 1 shows the estimation results obtained for 100 simulations of samples of size $N = 5000$. To compute $\hat{\theta}$ we use equation (14), where $\{\hat{a}_k\}$ are given by (9). The left panel on the figure shows that large values of $1/\lambda_N$ are associated with large median biases. This emphasizes the need to penalize the estimation. On the other hand, very small values of $1/\lambda_N$ penalize too much, and there is also some bias. Intermediate values of $1/\lambda_N$ seem to be best. We computed the mean squared error of $\hat{\theta}$ across simulations, and chose the minimum value $1/\lambda_N = 3.3 * 10^3$ to estimate the regression function.

The right panel on the figure shows the estimate of the regression function, computed using (13). Note that this estimator does not assume knowledge of the parametric form of $\mathbb{E}(Y|X^* = x^*)$. The estimate is quite close (in median) to the true regression function. We note some bias for large values of $|x^*|$. There are several possible reasons for the bias. Maybe the dictionary of functions that we use is not large enough (we use $25^2 = 625$ functions). This problem seems to be inherent to latent variables models, which are high dimensional and thus require to use a very large dictionary of functions.

A second possible reason for the bias is that SPADES is based on Lasso, and it has been shown that post-Lasso may improve over Lasso in finite samples (Belloni and Chernozhukov, 2010a). Nevertheless, the results we obtain are encouraging, and suggest that SPADES may be successfully applied to simple LVM.

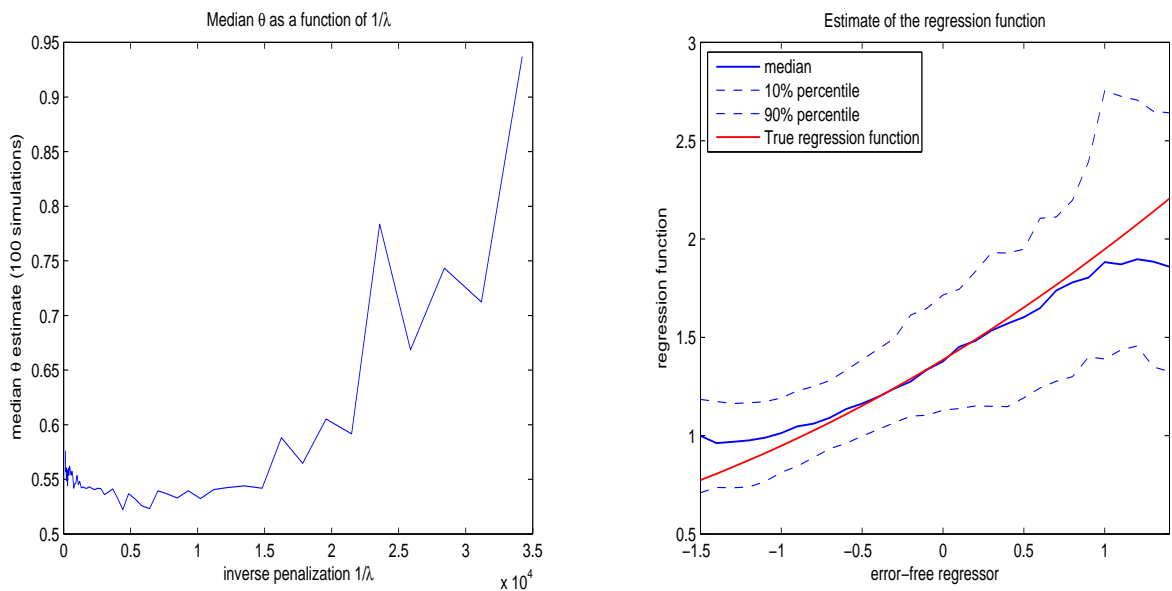
4 Conclusion and directions for future research

Latent variables models are high-dimensional, due to the presence of unknown distribution functions. Penalization methods provide interesting candidate estimators in these models. In particular, ℓ^1 penalization tends to select parsimonious, easily interpretable models. Another attractive feature of this general approach is that it can accommodate a very rich model space, using a large dictionary of functions. To make this procedure operational, we have used the penalized least squares minimization approach of Bunea *et al.* (2010), which we have applied to a simple measurement error model.

Recent developments on ℓ^1 penalization make Lasso-based approaches attractive from several points of view. From a computational perspective, fast algorithms are now available to compute the full path of Lasso solutions. From a statistical point of view, it has been shown that Lasso satisfies sparsity oracle inequalities (Bickel *et al.*, 2008) and, in the context of density estimation, that SPADES satisfies optimality properties (Bunea *et al.*, 2010). Progress has also been made to characterize the asymptotic behavior of Lasso estimators (Knight and Fu, 2000, Huang *et al.*, 2008).

Extending the approach to more general latent variables models raises a number of issues. In applications, unknown structural parameters may be present along with latent distributions. This situation will arise in the measurement error model of section 3.2 if $f_{Y|X^*}$ depends on a parameter vector θ . In this case, one may want to use SPADES to jointly estimate θ and the density of X^* . Another issue that arises frequently in applications of

Figure 1: Measurement error model ($N = 5000$, true $\theta = 1/2$)



Note: The left panel reports the median $\hat{\theta}$ obtained from (14), across 100 simulations, as a function of the inverse penalization parameter $1/\lambda_N$. The right panel shows the median, 10% and 90% pointwise percentiles of $\hat{\theta}$ computed by (13), for the value $1/\lambda_N = 3.3 * 10^3$.

LVM is the presence of conditioning covariates. When conditioned on continuous covariates, the approach that we have proposed faces a challenging curse of dimensionality.

Repeated measurements. We end this note by mentioning another example, which we borrow from Susanne Schennach’s paper that appears in this volume. The example is the simple measurement error model of Section 2, where now we wish to relax the assumption that $f_{X|X^*}$ is known. Instead, we suppose that we have two independent repeated measures of X^* :

$$\begin{cases} X &= X^* + U \\ \tilde{X} &= X^* + W. \end{cases}, \quad \text{where } U, W, X^* \text{ are mutually independent.}$$

Kotlarski (1967) shows that, under weak conditions, the distributions of X^* , U and W are nonparametrically identified in this model. Here we wish to estimate those distributions.

We start by noting the following identity:

$$f_{X,\tilde{X}}(x,\tilde{x}) = \int f_U(x-x^*) f_W(\tilde{x}-x^*) f_{X^*}(x^*) dx^*. \quad (15)$$

Following a similar approach as before, let $\{\Psi_k\}$, $\{\zeta_\ell\}$, and $\{\nu_m\}$ be three dictionaries of functions, with K , L , and M elements, respectively. Let us specify:

$$f_{X^*} = \sum_{k=1}^K a_k \Psi_k, \quad f_U = \sum_{\ell=1}^L b_\ell \zeta_\ell, \quad \text{and} \quad f_W = \sum_{m=1}^M c_m \nu_m.$$

Then, (15) yields:

$$\begin{aligned} f_{X,\tilde{X}}(x,\tilde{x}) &= \sum_{k=1}^K \sum_{\ell=1}^L \sum_{m=1}^M a_k b_\ell c_m \int \zeta_\ell(x-x^*) \nu_m(\tilde{x}-x^*) \Psi_k(x^*) dx^* \\ &= \sum_{k=1}^K \sum_{\ell=1}^L \sum_{m=1}^M a_k b_\ell c_m \varphi_{k,\ell,m}(x,\tilde{x}), \end{aligned}$$

where $\varphi_{k,\ell,m}(x,\tilde{x}) = \int \zeta_\ell(x-x^*) \nu_m(\tilde{x}-x^*) \Psi_k(x^*) dx^*$ are known functions.

An ℓ^1 penalized estimator of $\{a_k\}$, $\{b_\ell\}$, and $\{c_m\}$ is:

$$\begin{aligned} \{\hat{a}_k, \hat{b}_\ell, \hat{c}_m\} &= \underset{\{a_k, b_\ell, c_m\}}{\operatorname{argmin}} - \frac{2}{N} \sum_{i=1}^N \sum_{k=1}^K \sum_{\ell=1}^L \sum_{m=1}^M a_k b_\ell c_m \varphi_{k,\ell,m}(x_i, \tilde{x}_i) \\ &\quad + \left\| \sum_{k=1}^K \sum_{\ell=1}^L \sum_{m=1}^M a_k b_\ell c_m \varphi_{k,\ell,m} \right\|_2^2 \\ &\quad + \lambda_N \left(\sum_{k=1}^K \|\Psi_k\|_\infty |a_k| + \sum_{\ell=1}^L \|\zeta_\ell\|_\infty |b_\ell| + \sum_{m=1}^M \|\nu_m\|_\infty |c_m| \right). \end{aligned}$$

This estimation problem differs from SPADES as the unpenalized objective function is not quadratic (it may actually be non-convex). We are not aware of algorithms that solve this type of polynomial programming problems efficiently. It seems of interest to generalize the approach that we have proposed to deal with this type of situations.

References

- [1] Belloni, A., and V. Chernozhukov (2010a): “Post- ℓ^1 -Penalized Estimators in High-Dimensional Linear Regression Models,” CEMMAP Working Paper 1310.
- [2] Belloni, A., and V. Chernozhukov (2010b): “ ℓ^1 -Penalized Quantile Regression in High-Dimensional Sparse Models,” forthcoming *Annals of Statistics*.
- [3] Belloni, A., Chen, D., Chernozhukov, V., and C. Hansen (2010): “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain,” unpublished manuscript.
- [4] Bester, A., and C. Hansen (2007): “Flexible Correlated Random Effects Estimation in Panel Models with Unobserved Heterogeneity,” unpublished manuscript.
- [5] Bickel, P. J., Ritov, Y., and A. B. Tsybakov (2009): “Simultaneous Analysis of Lasso and Dantzig Selector,” *Ann. Statist.*, 37, 1705–1732.
- [6] Birgé, L., and P. Massart (1997): “From Model Selection to Adaptive Estimation,” In *Festschrift for Lucien LeCam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen, and G. Yang, eds., 55–87. Springer, New York.
- [7] Bonhomme, S., and J. M. Robin (2010): “Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics,” *Review of Economic Studies*, 77(2), 491–533.
- [8] Bunea, F., Tsybakov, A., Wegkamp M., and A. Barbu (2010): “SPADES and Mixture Models,” *Ann. Stat.*, 38(4), 2525–2558.
- [9] Chen, X. (2007): “Large Sample Sieve Estimation of Semi-nonparametric Models,” chapter 76 in *Handbook of Econometrics*, vol. 6B, 2007, eds. J.J. Heckman and E.E. Leamer, North-Holland.
- [10] Chen, X., Hong, H., and E. Tamer (2005): “Measurement Error Models with Auxiliary Data,” *Review of Economic Studies*, 72, 343–366.
- [11] Efron, B., Hastie, T., Johnstone, I., and R. Tibshirani (2004): “Least Angle Regression,” *Ann. Statist.*, 32, 407–499.
- [12] Evdokimov, K. (2010): “Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity,” unpublished manuscript.
- [13] Friedman, J., Hastie, T., and R. Tibshirani (2010): “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software*, 33(1), 1–22. Matlab code available at: <http://www-stat.stanford.edu/~tibs/lasso.html>
- [14] Hastie, T., and M. Y. Park (2007): “ ℓ^1 -Regularization Path Algorithm for Generalized Linear Models,” *J. R. Statist. Soc. B*, 69, 659–677.
- [15] Horowitz, J. L., and M. Markatou (1996): “Semiparametric Estimation of Regression Models for Panel Data”, *Review of Economic Studies*, 63, 145–168.

- [16] Huang, J., Horowitz, J. L., and S. G. Ma (2008): “Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models,” *Annals of Statistics*, 36, 587–613.
- [17] Hu, Y., and S.M. Schennach (2008): “Instrumental Variable Treatment of Nonclassical Measurement Error Models,” *Econometrica*, 76(1), 195–216.
- [18] Hu, Y., and M. Shum (2009): “Nonparametric Identification of Dynamic Models with Unobserved State Variables,” unpublished manuscript.
- [19] Knight, K., and W. Fu (2000): “Asymptotics for Lasso-Type Estimators,” *Annals of Statistics*, 28(5), 1356–1378.
- [20] Kotlarski, I. (1967): “On Characterizing the Gamma and Normal Distribution,” *Pacific Journal of Mathematics*, 20, 69–76.
- [21] Shen, X. (1997): “On Methods of Sieves and Penalization,” *The Annals of Statistics*, 25, 2555–2591.
- [22] Székely, G.J., and C.R. Rao (2000): “Identifiability of Distributions of Independent Random Variables by Linear Combinations and Moments,” *Sankhyä*, 62, 193–202.
- [23] Tibshirani, R. (1996): “Regression Shrinkage and Selection via the Lasso,” *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.