PENALIZED LEAST SQUARES MODEL AVERAGING

Erkki P. Liski[1] *and Antti Liski*[2]

[1]School of Information Sciences, University of Tampere,
FIN-33014 Tampere, FINLAND, Erkki.Liski@uta.fi
[2]Institute of Signal Processing, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND, Antti.Liski@tut.fi

ABSTRACT

In model selection one attempts to use the data to find a single "winning" model, whereas with model averaging (MA) one seeks a smooth compromise across a set of competing models. Most existing MA methods are based on estimation of single model weights using some appropriate criterion. The problem of selecting the best subset or subsets of predictor variables is a common challenge for a regression analyst. The number of candidate models may become huge and any approach based on estimation of all single weights may become computationally infeasible. Our approach is to convert estimation of model weights into estimation of shrinkage factors with trivial computational burden. We define the class of shrinkage estimators in view of MA and show that the estimators can be constructed using penalized least squares (LS) estimation by putting appropriate restrictions on the penalty function. The relationship between shrinkage and parameter penalization provides tools to build up computationally efficient MA estimators which are easy to implement into practice.

1. THE MODEL

Our framework is the linear model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n\times p$ and $n\times m$ matrices of nonrandom regressors, $(\mathbf{X}, \mathbf{Z})$ is assumed to be of full column-rank $p + m < n$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $p \times 1$ and $m \times 1$ vectors of unknown parameters. Our interest is in the effect of $\mathbf{X}$ on $\boldsymbol{y}$, that is, we want to estimate $\boldsymbol{\beta}$ while the role of $\mathbf{Z}$ is to improve the estimation of $\boldsymbol{\beta}$.

We will work with the canonical form of the model (1), where z-variables are orthogonalized by writing the systematic part of the model (1) as

$$\begin{aligned} \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} &= \mathbf{X}\boldsymbol{\alpha} + \mathbf{M}\mathbf{Z}\boldsymbol{\gamma} \\ &= \mathbf{X}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\theta}, \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}$,

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} \quad \text{and} \quad \mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{3}$$

are symmetric idempotent matrices. Since $(\mathbf{M}\mathbf{Z})'\mathbf{M}\mathbf{Z} = \mathbf{Z}'\mathbf{M}\mathbf{Z}$ is positive definite [15], then there exists a nonsingular matrix $\mathbf{C}$ such that [9]

$$\mathbf{C}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{C} = (\mathbf{M}\mathbf{Z}\mathbf{C})'(\mathbf{M}\mathbf{Z}\mathbf{C}) = \mathbf{U}'\mathbf{U} = \mathbf{I}_m. \tag{4}$$

In (4) $\mathbf{U} = \mathbf{M}\mathbf{Z}\mathbf{C}$ denotes the matrix of orthogonal canonical auxiliary regressors. Introducing the canonical auxiliary parameters $\boldsymbol{\theta} = \mathbf{C}^{-1}\boldsymbol{\gamma}$ we can write in (2)

$$\mathbf{M}\mathbf{Z}\boldsymbol{\gamma} = \mathbf{M}\mathbf{Z}\mathbf{C}\mathbf{C}^{-1}\boldsymbol{\gamma} = \mathbf{U}\boldsymbol{\theta}.$$

2. MODEL AVERAGING

A least squares MA estimator for $\boldsymbol{\beta}$ takes the form

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \sum_{i=0}^{M} \lambda_i \hat{\boldsymbol{\beta}}_i = \sum_{i=0}^{M} \lambda_i (\hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\mathbf{W}_i\hat{\boldsymbol{\theta}}) \\ &= \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\mathbf{W}\hat{\boldsymbol{\theta}}, \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$, $\mathbf{W} = \sum_{i=0}^{M} \lambda_i \mathbf{W}_i$ and $\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{C}$. The weights

$$\lambda_i = \lambda_i(\mathbf{M}\boldsymbol{y}) \geq 0, \quad i = 0, 1, \ldots, M,$$

are assumed to depend on the least squares residuals $\mathbf{M}\boldsymbol{y}$ and $\sum_{i=0}^{M} \lambda_i = 1$. Note especially that $\hat{\boldsymbol{\theta}}$ is a function of $\mathbf{M}\boldsymbol{y}$. The selection matrices $\mathbf{W}_i$, $0 \leq i \leq M$ are nonrandom $m \times m$ diagonal matrices with diagonal elements 0 or 1 whereas $\mathbf{W}$ is a random $m \times m$ diagonal matrix with diagonal elements

$$\boldsymbol{w} = (w_1, \ldots, w_m)', \quad 0 \leq w_i \leq 1,\ i = 1, \ldots, m.$$

The equivalence theorem of Danilov and Magnus [3] provides a useful representation for the expectation, variance and MSE of the estimator $\tilde{\boldsymbol{\beta}}$ given in (5). The theorem was proved under the assumptions that the disturbances $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$. By the theorem

$$\begin{aligned} MSE(\tilde{\boldsymbol{\beta}}) &= E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}[MSE(\mathbf{W}\hat{\boldsymbol{\theta}})]\mathbf{Q}'. \end{aligned}$$

The quality of $\tilde{\boldsymbol{\beta}}$ essentially depends on statistical properties of the shrinkage estimator $\mathbf{W}\hat{\boldsymbol{\theta}}$ and hence the relatively simple estimator $\mathbf{W}\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ characterizes the important features of the more complicated estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. It can be shown (Hansen [8]) that a least squares MA estimator like (5) can achieve lower MSE than any individual LS estimator.

3. PENALIZED LS AND SRINKAGE

We introduce a set $\mathcal{S}$ of shrinkage estimators for $\boldsymbol{\beta}$ and characterize them by using penalized least squares technique. Then we derive the efficiency bound for the shrinkage estimators with respect to MSE (mean squared error) when observations follow the normal distribution. Our aim is to find estimators whose MSE is uniformly as close to the efficiency bound as possible. It turns out that many interesting known estimators, like for example the soft and firm thresholding estimators, non-negative garrote [2] and the SCAD (smoothly clipped absolute deviation, [6]) estimators belong to this shrinkage class $\mathcal{S}$. On the other hand, for example the hard thresholding rule (pre testing) and the ridge estimator do not belong to $\mathcal{S}$.

Fitting the orthogonalized model (2) can be considered as a two-step least squares procedure [15]. The first step is to calculate $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ and to replace $\boldsymbol{y}$ by $\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0 = \mathbf{M}\boldsymbol{y}$, where $\mathbf{M}$ is defined in (3). Then denote $\boldsymbol{z} = \mathbf{U}'\boldsymbol{y}$, and note from the definition of $\mathbf{U}$ in (4) the equality $\mathbf{U}'\mathbf{M} = \mathbf{U}'$. Then the model (2) takes the form

$$\boldsymbol{z} = \boldsymbol{\theta} + \mathbf{U}'\boldsymbol{\varepsilon}, \qquad \mathbf{U}'\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_m). \tag{6}$$

The second step is to estimate $\boldsymbol{\theta}$ from the model (6).

Magnus et al. [13] estimated the weights $0 \le w_i \le 1,\ i = 1, \ldots, m$ in (5) using a Bayesian technique, and decided on to advocate the Laplace estimator which is of a shrinkage type. Such estimators are computationally superior to estimators that require estimation of every single model weight λ_i in (5). We are now ready to define the important class $\mathcal{S}$ of shrinkage estimators for θ which we call simply shrinkage estimators.

Definition A real valued estimator δ of θ is a shrinkage estimator if the following four conditions hold:

(a) $0 \le \delta(\hat{\theta}) \le \hat{\theta}$ for $\hat{\theta} \ge 0$,

(b) $\delta(-\hat{\theta}) = -\delta(\hat{\theta})$,

(c) $\delta(\hat{\theta})/\hat{\theta}$ is nondecreasing on $[0, \infty)$ and

(d) $\delta(\hat{\theta})$ is continuous,

where $\hat{\theta}$ is the LS estimator of θ.

In estimation of $\boldsymbol{\theta}$ we will use the penalized LS technique. If the penalty function satisfies proper regularity conditions, the penalized LS yields a solution which is a shrinkage estimator of $\boldsymbol{\theta}$. In this approach we choose a suitable penalty function in order to get a shinkage estimator with good risk properties. The penalized least squares estimate (PenLS) of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ is the minimizer of

$$\frac{1}{2}\sum_{i=1}^{m}(z_i - \theta_i)^2 + \sum_{i=1}^{m} p_\lambda(|\theta_i|), \tag{7}$$

where $\lambda > 0$. It is assumed that the penalty function $p_\lambda(\cdot)$ is

$$\begin{aligned} &\text{(i)} && \text{nonnegative,} \\ &\text{(ii)} && \text{nondecreasing and} \\ &\text{(iii)} && \text{differentiable on } [0, \infty). \end{aligned} \tag{8}$$

Minimization of (7) is equivalent to minimization componentwise. Thus we may simply minimize

$$l(\theta) = \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|) \tag{9}$$

with respect to θ.

Example There are close connections between the PenLS and variable selection or the PenLS and ridge regression, for example. Taking the L_2 penalty $p_\lambda(|\theta|) = \frac{\lambda}{2}|\theta|^2$ yields the ridge estimator

$$\check{\theta}_R = \frac{1}{1+\rho}z,$$

where $\rho > 0$ depends on λ. The hard thresholding penalty function

$$p_\lambda(|\theta|) = \lambda^2 - \frac{1}{2}(|\theta| - \lambda)^2(\mathrm{I}(|\theta| < \lambda)$$

yields the hard thresholding rule

$$\check{\theta}_H = z\,\{\mathrm{I}(|\mathrm{z}| > \lambda)\}, \tag{10}$$

where $\mathrm{I}(\cdot)$ is the indicator function. Then the minimizer of the expression (7) is $z_j\{\mathrm{I}(|\theta_\mathrm{j}| > \lambda)\},\ \mathrm{j} = 1, \ldots, \mathrm{m}$, and it coincides with the best subset selection for orthonormal designs. In statistics (see e.g. Morris et al. [14]) and in econometrics (see, e.g. Judge *et al.* [10]), the hard thresholding rule is traditionally called the pretest estimator.

The following theorem gives sufficient conditions for the PenLS estimate $\check{\theta}$ of θ to be a shrinkage estimator. Further, the theorem provides the lower bound of the mean squared error

$$MSE(\theta, \check{\theta}) = E[\check{\theta}(z) - \theta]^2. \tag{11}$$

This lower bound is called the *efficiency bound*.

Theorem 3.1. *We assume that the penalty function $p_\lambda(\cdot)$ satisfies the assumptions* (8)*. We make two assertions.*

(i) If the three conditions hold

(1) the function $-\theta - p'_\lambda(\theta)$ is strictly unimodal on $[0, \infty)$,

(2) $p'_\lambda(\cdot)$ is continuous and nonincreasing on $[0, \infty)$, and

(3) $\min_\theta\{|\theta| + p'_\lambda(|\theta|)\} = p'_\lambda(0)$,

then the PenLS estimate $\check{\theta}$ of θ belongs to the shrinkage family $\mathcal{S}$.

(ii) If the conditions of the assertion (i) hold and z follows the normal distribution $\mathrm{N}(0,\sigma^2)$, where σ^2 is known, the efficiency bound of $\check{\theta}$ is

$$\inf_{\check{\theta}\in\mathcal{S}} MSE(\theta,\check{\theta}) = \frac{\theta^2}{1+\theta^2}. \tag{12}$$

Note that the pretest estimator $\check{\theta}_H$ given in (10) is not continuous, and hence it does not belong to the class of shrinkage estimators $\mathcal{S}$. Magnus [11] demonstrates a number of undesiderable properties of the pretest estimator. It is inadmissible and there is a range of values for which the MSE of $\check{\theta}_H$ is greater than the MSE of both the least squares estimator $\hat{\theta}(z) = z$ and the null estimator $\hat{\theta}(z) \equiv 0$. The traditional pretest at the usual 5% level of significance results in an estimator that is close to having worst possible performance with respect to the MSE criterion in the neighborhood of the value $|\theta/\sigma| = 1$ which was shown to be of crucial importance.

Example The L_q penalty $p_\lambda(|\theta|) = \lambda\,|\theta|^q,\ q \geq 0$ results in a bridge regression [7]. The derivative $p'_\lambda(\cdot)$ of the L_q penalty is nonincreasing on $[0,\infty)$ only when $q \leq 1$ and the solution is continuous only when $q \geq 1$. Therefore, only L_1 penalty in this family yields a shrinkage estimator. This estimator is the soft thresholding rule, proposed by Donoho and Johnstone [4],

$$\check{\theta}_S = \mathrm{sgn}(z)(|z|-\lambda)_+, \tag{13}$$

where z_+ is shorthand for $\max\{z,0\}$. LASSO [16] is the PenLS estimate with the L_1 penalty in the general least squares and likelihood settings.

If the PenLS estimators satisfy the conditions of Theorem 3.1, the efficiency bound is known and the *regret* of $\check{\theta}(z)$ can be defined as

$$r(\theta,\check{\theta}) = MSE(\theta,\check{\theta}) - \frac{\theta^2}{1+\theta^2}.$$

We wish to find an estimator with the desirable property that its MSE is uniformly close to the infeasible efficiency bound. In theoretical considerations σ^2 is assumed to be known, and hence we can always consider the variable z/σ. Then the expectation E is simply taken with respect to the $N(\theta,1)$ distribution (cf. Figure 1), and comparison of estimators risk performance is done under this assumption. In practical applications we replace the unknown σ^2 with s^2, the estimate of σ^2 in the unrestricted model. Danilov [3] demonstrated that effects of estimating σ^2 are small in case of Laplace estimator. We expect the approximation to be accurate for other shrinkage estimators too, although more work is needed to clarify this issue.

3.1. Good PenLS shrinkage estimators

In this subsection we consider properties of two well known PenLS estimators which are shrinkage estimators. Bruce and Gao [1] compared the hard and soft thresholding rules and showed that the hard thresholding rule tends to have bigger variance than the soft thresholding rule whereas soft thresholding tends to have bigger bias. To remedy the drawbacks of hard and soft thresholding, Fan and Li [6] suggested using continuous differentiable penalty function defined by

$$p'_\lambda(|\theta|) = \lambda\left\{\mathrm{I}(|\theta|\leq\lambda) + \frac{(\mathrm{a}\lambda-|\theta|)_+}{(\mathrm{a}-1)\lambda}\mathrm{I}(|\theta|>\lambda)\right\} \tag{14}$$

for some $a > 2$ and $\lambda > 0$. If the penalty function in (7) is constant, i.e. $p'(|\theta|) = 0$, then the PenLS takes the form $\check{\theta}(z) \equiv z$ which is unbiased. Since the SCAD penalty $p'_\lambda(\theta) = 0$ for $\theta > a\lambda$, the resulting solution (Fan and Li [6])

$$\check{\theta}_{scad}(z) = \begin{cases} \mathrm{sgn}\,(z)(|z|-\lambda)_+, & \text{if } |z| \leq 2\,\lambda, \\ \frac{(a-1)z-\mathrm{sgn}\,(z)a\lambda}{(a-2)}, & \text{if } 2\lambda < |z| \leq a\lambda, \\ z, & \text{if } |z| > a\lambda \end{cases} \tag{15}$$

tends to be unbiased for large values of z. The estimator (15) can be viewed as a combination of soft thresholding for "small" $|z|$ and hard thresholding for "large" $|z|$, with a piecewise linear interpolation inbetween.

Breiman [2] applied the non-negative garrote rule

$$\check{\theta}_G(z) = \begin{cases} 0, & \text{if } |z| \leq \lambda, \\ z - \lambda^2/z, & \text{if } |z| > \lambda \end{cases} \tag{16}$$

to subset selection in regression to overcome the drawbacks of stepwise variable selection rule and ridge regression. It is straightforward to show that the soft thresholding (13), SCAD (15) and non-negative garrote (16) estimators belong to the shrinkage class $\mathcal{S}$ (cf. Definition). The ordinary LS (OLS) estimator $\hat{\theta}(z) \equiv z$ is a good candidate for large z, and hence we wish that for large z an estimator $\check{\theta}(z)$ is close to z in the sense that $z - \check{\theta}(z)$ converges to zero when $|z|$ increases. It can be readily seen that the estimators $\check{\theta}_{scad}$ and $\check{\theta}_G$ have this property. For the soft thresholding rule $z - \check{\theta}_S(z)$ converges to a positive constant, but not to zero.

3.2. The Laplace and Subbotin estimators

Magnus [12] addressed the question of finding an estimator of θ which is admissible, has bounded risk, has good risk performance around $\theta = 1$, and is optimal or near optimal in terms of minimax regret when $z \sim \mathrm{N}(\theta,1)$. The Laplace estimator

$$\hat{\theta}_L(z) = z - h(y)c \tag{17}$$

proved to be such an estimator, when $c = log\,2$ and $h(\cdot)$ is a given antisymmetric monotonically increasing function on $(-\infty,\infty)$ with $h(0) = 0$ and $h(\infty) = 1$. The Laplace estimator is the mean of the posterior distribution of $\theta|z$ when a Laplace prior for θ with $\mathrm{median}(\theta) = 0$ and $\mathrm{median}(\theta^2) = 1$ is assumed. In search of a prior which appropriately reflects the notion of ignorance, Einmahl et al. [5] arrived at the Subbotin prior that belongs to the

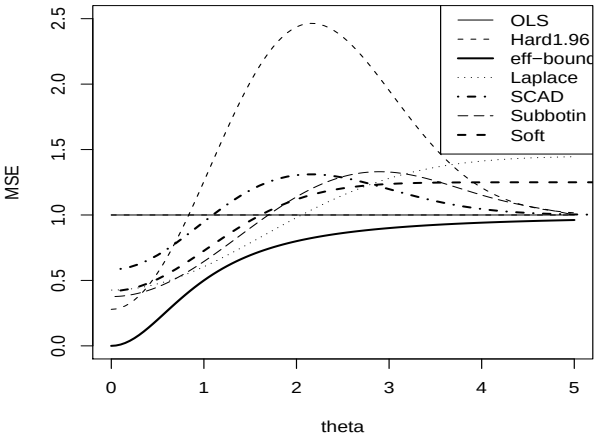

Figure 1. MSE of the OLS, the hard thresholding (10), Laplace (17), SCAD (15) , Subbotin, soft thresholding estimators (13) and the effieciency bound (12) for the shrinkage estimators $\mathcal{S}$.

class of reflected gamma densities. In practical applications they recommended the Subbotin prior

$$\pi(\theta) = \frac{c^2}{4} e^{-c|\theta|^{1/2}}$$

with $c = 1.6783$ which should stay close to the Laplace prior. .

4. CONCLUDING REMARKS

Many existing MA methods require estimation of every single model weight. For example, in regression analysis selection of the best subset from a set of m predictors, say, requires assessing 2^m models, and consequently the computational burden soon increases too heavy when m becomes large.

It turns out, that the quality of the least squares MA estimator (5) depends on the shrinkage estimator of the auxiliary parameter $\boldsymbol{\gamma}$. So, estimation of 2^m model weights is converted into estimation of m shrinkage factors with trivial computational burden. We define the class of shrinkage estimators in view of MA and show that these shrinkage estimators can be constructed by putting appropriate restrictions on the penalty function. Utilizing the relationship between shrinkage and parameter penalization, we are able to build up computationally efficient MA estimators which are easy to implement into practice. These estimators include some well known estimators, like the nonnegative garrote of Breiman [2], the lasso-type estimator of Tibshirani [16] and the SCAD estimator of Fan and Li [6]. In the simulation experiments we have assessed the quality of estimators in terms of estimated MSE's. In this competition the winners were the SCAD and nonnegative garrote but the Laplace estimator did almost as well. However, the results of the simulation study are not reported here.

5. REFERENCES

[1] Bruce, A. G. and Gao, H.-Y. (1996). Understanding WaveShrink: Variance and bias estimation. *Biometrica*, 83, 727–745.

[2] Breiman, L. (1995). Better subset regression using nonnegative garrote.. *Technometrics*, 37, 373–384.

[3] Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics,* 122, 2746.

[4] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425–456.

[5] Einmahl, J. H. J., Kumar, K. and Magnus J. R. (2011) Bayesian model averaging and the choice of prior. *CentER Discussion Paper*, No. 2011-003.

[6] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American statistical Association*, 96, 1348–1360.

[7] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109–148.

[8] Hansen, B. E. (2007). Least squares model averaging. *Econometrika*, 75, 1175–1189.

[9] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge, Cambridge University Press.

[10] Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H. and Lee, T. C.(1985). *The Theory and Practice of Econometrics*, Wiley, New York.

[11] Magnus, J. R. (1999). The traditional pretest estimator. Theory of Probability and Its Applications, 44, 293308.

[12] Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with a known variance. *Econometrics Journal,* 5, 225236.

[13] Magnus, J. R., Powell, O. and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometris*, 154, 139–153.

[14] Morris, C., Radhakrishnan, R. and Sclove, S. L. (1972). Nonoptimality of preliminary test estimators for the mean of a multivariate normals distribution. *Annals of Mathematical Statistics*, 43, 1481–1490.

[15] Seber, G. A. F. (1977). *Linear Regression Analysis*, New York, Wiley.

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, 1, 267–288.