# On the practice of rescaling covariates

By SYLVAIN SARDY*

Summary: Whether doing parametric or nonparametric regression with shrinkage, thresholding, penalized likelihood, Bayesian posterior estimators (e.g., *ridge regression*, *lasso*, *principal component regression*, *waveshrink* or *Markov random field*), it is common practice to rescale covariates by dividing by their respective standard errors $\rho$. The stated goal of this operation is to provide unitless covariates to compare like with like, especially when penalized likelihood or prior distributions are used. We contend that this vision is too simplistic. Instead, we propose to take into account a more essential component of the structure of the regression matrix by rescaling the covariates based on the diagonal elements of the covariance matrix $\Sigma$ of the maximum likelihood estimator. We illustrate the differences between the standard $\rho$- and proposed $\Sigma$-rescalings with various estimators and data sets.

**Key Words:** lasso, Markov random field, principal component regression, $\ell_\eta$ prior, ridge regression, wavelets.

*Address for correspondence: Université de Genève, Section de Mathématiques, 2-4 rue du Lièvre, 1211 Genève, Switzerland. Email: Sylvain.Sardy@math.unige.ch

Résumé: Que l'on utilise un modèle de régression paramétrique ou non-paramétrique, par rétrécissement, seuillage, vraisemblance pénalisée ou Bayesien (ex. régression ridge, lasso, régression en composantes principales, waveshrink, champ Markovien), il est commun de standardiser les variables explicatives en les divisant par leurs écarts types $\rho$ respectifs. Le but affiché de cette opération est de créer des variables sans unités pour pouvoir les comparer entre elles, en particulier quand l'estimateur est basé sur la vraisemblance pénalisée ou une distribution à priori. Nous attendons prouver que cette vision est trop simpliste. Nous proposons de plutôt considérer un élément plus essentiel de la matrice de régression en standardisant les variables explicatives à partir des éléments diagonaux de la matrice de covariance $\Sigma$ de l'estimateur du maximum de vraisemblance. Nous illustrons les différences entre la standardisation $\rho$ et la standarisation $\Sigma$ avec des estimateurs et des données variés.

**Mots clés:** champ markovien, distribution a priori $\ell_\eta$, lasso, ondelettes, régression en composantes principales, régression ridge.

# 1 Background

Linear models are extensively used in parametric and nonparametric regression. In parametric regression, $P$ covariates $\mathbf{X} = (x_1, \ldots, x_P)$ and an associated noisy response $Y$ are sampled $N$ times from their joint distribution. The regression problem consists in assessing the association $\mu(\mathbf{x})$ between covariates and response based on the data set $\{(Y_n, \mathbf{x}_n)\}_{n=1,\ldots,N}$. Parametric linear models assume $\mu(\mathbf{x}_n) = \alpha_0 + \sum_{p=1}^{P} \alpha_p x_{np}$, where $\alpha_0$ is the intercept coefficient and $\alpha_p$ are coefficients associated to the covariates. Additional covariates can be built from the original ones by means of simple transformations (e.g., $\log x$ or $x^p$); we assume here that the $P$ covariates include all the covariates the practitioner believes useful. Likewise expansion-based nonparametric models assume a low dimensional function $\mu(t)$, say univariate, can be well approximated by a linear combination $\alpha_0 + \sum_{p=1}^{P_N} \alpha_p \varphi_p(t)$ of a large number $P_N = O(N)$ of known basis functions $\varphi_p(t)$ such as splines or wavelets. Both parametric and nonparametric models write in vector notation as $\alpha_0 \mathbf{1} + X\boldsymbol{\alpha}$, where $X$ is either the matrix of covariates or the matrix of discretized basis functions, and $(\alpha_0, \boldsymbol{\alpha})$ is the unknown vector of coefficients.

A primary goal of regression is to estimate the coefficients $(\alpha_0, \boldsymbol{\alpha})$ to achieve good predictive performance. Sometimes, interpretability is also of interest, in which case most estimated coefficients should be set to zero to allow interpretation of the sign and magnitude of the remaining non-zero coefficients. For its simplicity and its availability in most softwares, the least squares estimates solution to

$$\min_{\alpha_0, \boldsymbol{\alpha}} \|\mathbf{y} - (\alpha_0 \mathbf{1} + X\boldsymbol{\alpha})\|_2^2 \tag{1}$$

is widely used. With the $\ell_2$ loss we implicitly assume that the noise is additive, but, replacing (1) by the deviance, our approach extends to generalized linear models (Nelder and Wedderburn, 1972; Wahba, 1990), as we illustrate with Poisson data in Section 4.4. Gauss-Markov theorem states the seemingly advantageous property that least squares is the best linear unbiased estimator when the errors are additive and uncorrelated, and have expectation zero and equal variances. The least squares

estimator has some drawbacks however. First, although unbiased, it can have a large variance and therefore a bad predictive performance when the regression matrix has nearly collinear columns. And when the columns of the regression matrix are linearly dependent (e.g., $P > N$), the least squares estimate is not even uniquely defined. Second, the fitted model is not interpretable since all entries of the least squares estimate are different from zero with probability one.

Regularization is intended to control variance, at the cost of introducing bias. A class of regularization methods adds a penalty to (1) and solves

$$\min_{\alpha_0,\boldsymbol{\alpha}} \|\mathbf{y} - (\alpha_0\mathbf{1} + X\boldsymbol{\alpha})\|_2^2 + \lambda\|B\boldsymbol{\alpha}\|_\eta^\eta, \tag{2}$$

where the hyperparameter $\lambda \geq 0$, the penalty $\|\boldsymbol{\mu}\|_\eta^\eta = \sum_{i=1}^{Q} |\mu_i|^\eta$ for $\eta \geq 0$ (here $Q$ is the length of the vector $\boldsymbol{\mu}$), the matrix $B$ is $Q \times P$, and the intercept coefficient $\alpha_0$ is usually not penalized. Equivalently, (2) can be seen as a negative log-posterior distribution, where the penalty is the negative log-prior distribution of the coefficients.

*Best subset variable selection* is the most common regularization method, and corresponds to (2) with $B = I$ and $\eta = 0$, where $\|\boldsymbol{\alpha}\|_0^0 = \sum_{p=1}^{P} |\alpha_p|^0$ counts the number of non-zero values of $\boldsymbol{\alpha}$ (with the convention that $0^0 = 0$); it is nonlinear and leads to an interpretable model since many coefficients are set to zero. *Ridge regression* (Hoerl and Kennard, 1970), with $B = I$ and $\eta = 2$, is linear, but does not lead to an interpretable model; Jensen and Ramirez (2008) give a recent account on ridge regression as a constrained optimization problem. *Lasso* (Tibshirani, 1996), with $B = I$ and $\eta = 1$, is nonlinear and interpretable. *Bridge regression* (Fu, 1998) links *ridge regression* to *lasso*. *Support vector machine* (Vapnik, 1995) with $B = I$ and $\eta = 2$ uses the $\epsilon$-insensitive loss $l(r) = (|r| - \epsilon) \cdot 1(|t| \geq \epsilon)$ in place of $\ell_2$ loss; it is nonlinear and uninterpretable. *Smoothing splines* (Wahba, 1990), with the matrix of first order finite differences for $B$ and $\eta = 2$, is linear. *Waveshrink* (Donoho and Johnstone, 1994), with $B = I$ and $\eta = 1$, is nonlinear and estimated wavelet coefficients are thresholded to zero. Other regularization methods exist that cannot be written in the form (2), for instance *partial least squares* (Wold,

1966), *least angle regression* (Efron, Hastie, Johnstone, and Tibshirani, 2004), and *boosting* which bears similarities with *lasso* (see Buehlmann and Hothorn (2007) for a review).

The aim of this paper is to show that rescaling the matrix $X$ should be considered and done properly before employing many of these estimators.

# 2 Rescaling

For ridge regression, Marquardt and Snee (1975) advocated rescaling the columns of the regression matrix, which is now a widespread practice. Rescaling is concerned with affine transformations (i.e., average centering and standardization) applied to the columns of X. Its motivation is not always clear however, as the debate following Smith and Campbell (1980)'s discussion paper shows. We review their standardization in Section 2.2.1 and present an alternative standardization in Section 2.2.2.

## 2.1 Average centering

The intercept is a special regressor that is not considered as a removable covariate while performing variable selection, so $\alpha_0$ is not penalized in (2). The practice of 'average-centering', or 'mean-centering' (Marquardt and Snee, 1975), is intended to separate the estimation of $\boldsymbol{\alpha}$ from that of $\alpha_0$. With the Gram-Schmidt operation of subtracting the average $\bar{y}$ from the response vector $\mathbf{y}$ (i.e., $\mathbf{y}_\bullet = \mathbf{y} - \bar{y}\mathbf{1}$), and the average $\bar{x}_p$ of the $p$th observed covariates $\mathbf{x}_p$ for $p = 1, \ldots, P$, the average-centered matrix $X_\bullet = (I - \frac{1}{N}J)X$, where $J$ is a square matrix of ones, is orthogonal to the constant vector. Hence the solution $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}})_\lambda$ of (2) is broken into two steps: first calculate $\hat{\boldsymbol{\alpha}}_\lambda$ as the solution to $\min_{\boldsymbol{\alpha}} \|\mathbf{y}_\bullet - X_\bullet\boldsymbol{\alpha}\|_2^2 + \lambda\|B\boldsymbol{\alpha}\|_\eta^\eta$, then calculate $\hat{\alpha}_{0,\lambda} = \bar{y} - \frac{1}{N}\mathbf{1}'X\hat{\boldsymbol{\alpha}}_\lambda$.

## 2.2 Standardization

Standardization of the columns of $X_\bullet$ amounts to right-multiply by a diagonal matrix $D$. For regularized estimators like (2) for instance, this operation is equivalent

to assuming an anisotropic prior since

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - X_\bullet D\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_\eta^\eta = \min_{\boldsymbol{\alpha}=D\boldsymbol{\beta}} \|\mathbf{y} - X_\bullet\boldsymbol{\alpha}\|_2^2 + \lambda\|D^{-1}\boldsymbol{\alpha}\|_\eta^\eta. \tag{3}$$

So standardization has an impact on the prior, and vice versa.

### 2.2.1 Classical $\rho$-standardization

The motivation for standardizing the columns of the regression matrix has led to controversy. Marquardt (1980) alleges interpretability of the estimated coefficients and a reason linked to the prior distribution of the coefficients. In a nutshell, the claim is first that if the covariates are standardized to be unitless (no apples with oranges), then the estimated coefficients are interpretable in the sense that they can be compared and removed from the model based on their relative sizes. Second, since the prior is typically isotropic (e.g., $\ell_\eta$ prior assumes same variance between independent coefficients), the covariates should be in the same unit or unitless. Take *lasso*'s $\ell_1$ equivalent prior for instance, the penalty is identical for each coefficient $\alpha_p$ in $+\lambda\sum_{p=1}^P |\alpha_p|$, which is wrong if the corresponding covariates are expressed in different units. For those reasons, Hoerl and Kennard (1970) and Marquardt and Snee (1975) propose to work with the correlation form of $X$, which has the following property.

**Property 1**: Let $\rho_p = \{\sum_{n=1}^N (x_{pn} - \bar{x}_p)^2\}^{1/2}$ for $p = 1, \ldots, P$ and $D_\rho = \text{diag}(\rho_1^{-1}, \ldots, \rho_P^{-1})$. Using $\rho$-standardization $X_{\bullet\rho} = X_\bullet D_\rho$ and letting $\hat{\boldsymbol{\beta}}_{\lambda,\bullet\rho}$ be the solution to

$$\min_{\boldsymbol{\beta}} \|\mathbf{y}_\bullet - X_{\bullet\rho}\boldsymbol{\beta}\|_2^2 + \lambda\|B\boldsymbol{\beta}\|_\eta^\eta,$$

then the estimated coefficients are scale invariant, for all $\lambda \geq 0$. The estimated coefficients of the original scale are then $\hat{\boldsymbol{\alpha}}_{\lambda,\rho} = D_\rho\hat{\boldsymbol{\beta}}_{\lambda,\bullet\rho}$. Note that $\hat{\boldsymbol{\alpha}}_{\lambda,\rho}$ obtained after rescaling is not the solution $\hat{\boldsymbol{\alpha}}_\lambda$ to (2).

**Proof**: First note that $\rho_p \neq 0$ for all $p$, otherwise the $p$th covariate can be integrated into the intercept. Let $X_\bullet$ be the original mean-centered matrix,

and let $X_\bullet^{(W)} = X_\bullet W$ reflect a change of units with the diagonal matrix $W = \mathrm{diag}(w_1, \ldots, w_P)$, e.g., $w_1 = 1000$ if the unit of the first covariate is changed from kilometers to meters. Letting $D_\rho$ be the rescaling matrix for $X_\bullet$, the rescaling matrix for $X_\bullet^{(W)}$ is $W^{-1}D_\rho$. The estimation is scale-invariant since $X_{\bullet\rho}^{(W)} = X_\bullet^{(W)} W^{-1} D_\rho = X_\bullet W W^{-1} D_\rho = X_{\bullet\rho}$. □

$\rho$-standardization should not be applied blindly, for instance if one believes that covariates expressed in the same units do not require standardization. Certain estimators are also invariant to standardization by nature, like best subset variable selection ($\eta = 0$). For instance, the multivariate adaptive regression splines (MARS) (Friedman, 1991) and TURBO (Friedman and Silverman, 1989) nonparametric estimators do not require rescaling of the splines used because the selection strategy is (forward–backward) stepwise. For those estimators, the advantage of $\rho$-standardizing, if any, is numerical.

### 2.2.2 $\Sigma$-standardization

All estimators of the form (2) require standardization as soon as $\lambda \neq 0$ and $\eta \neq 0$ due to the isotropic nature of the $\ell_\eta$ penalty. $\rho$-standardization attempts to put all covariates on an equal footing before estimation. But we contend that $\rho$-standardization does not operate properly. Take the least squares coefficients when $\lambda \to 0$: their relative significance is not determined by simple comparison of their magnitudes because they do not have the same variance, even if $X$ has been $\rho$-standardized. The least squares estimate must be rescaled by dividing by a measure of uncertainty; for instance, the $t$-statistics are unitless statistics obtained by division of each least squares coefficient by its respective standard error taken from the diagonal elements of the covariance matrix $\Sigma = (X_\bullet' X_\bullet)^{-1}$.

When a regularization method like (2) is employed, the heteroscedasticity of the least squares coefficients induced by the structure of the regression matrix $X$ must be reflected, as shown in (3), either in the weighting of the penalty or in the rescaling of the columns of $X$. We propose the following rescaling.

**Proposition 1**: Let $\hat{\boldsymbol{\alpha}}^{\mathrm{LS}}$ be the least squares estimate solution to (1) with covariance matrix proportional to $\Sigma = (X_\bullet' X_\bullet)^{-1}$ (take the Moore-Penrose generalized inverse if necessary) and let $D_\Sigma^2 = \mathrm{diag}(\Sigma)$. The covariance-based $\Sigma$-standardization is defined by $X_{\bullet\Sigma} = X_\bullet D_\Sigma$.

Besides being invariant to a change of scale in the covariates like $\rho$-standardization, $\Sigma$-standardization has the advantage that the least squares estimate $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{\bullet\Sigma}$ solution to (3) when $D = D_\Sigma$ and $\lambda \to 0$ has a covariance matrix with constant diagonal since $\mathrm{diag}((X_{\bullet\Sigma}' X_{\bullet\Sigma})^{-1}) = D_\Sigma^{-2}\mathrm{diag}((X_\bullet' X_\bullet)^{-1}) = I$ by definition of $D_\Sigma$. An estimator using an isometric prior of the form (2) can be employed with $X$ rescaled as $X_{\bullet\Sigma}$.

# 3 Estimators revisited with $\Sigma$-standardization

The goal of this section is to illustrate the heuristics of $\Sigma$-standardization with a few revealing examples.

## 3.1 Markov random field smoothing

Besag (1986) and Geman and Geman (1984) developed a nonparametric regression method based on a Markov random field prior for the function $\mu(t)$. Assuming a Gaussian ($\eta = 2$) or Laplace (Sardy and Tseng, 2004, $\eta = 1$) first order prior, the maximum a posteriori estimator solves

$$\min_{\boldsymbol{\mu}} \|\mathbf{y} - \boldsymbol{\mu}\|_2^2 + \lambda \sum_{n=1}^{N-1} |\mu_{n+1} - \mu_n|^\eta, \tag{4}$$

which belongs to the class of regularizations considered in (2) with $\boldsymbol{\alpha} = \boldsymbol{\mu}$, $X = I$, $\alpha_0 = 0$ and $B$ the sparse $(N-1) \times N$ matrix with $B_{n,n} = -B_{n,n+1} = -1$ for $n = 1, \ldots, N-1$ and zero otherwise. By a change of variable, this estimator can equivalently be written as an expansion-based estimator on Heaviside basis functions, i.e., $\boldsymbol{\mu} = \alpha_0 \mathbf{1} + X\boldsymbol{\alpha}$ with $(\alpha_0, \boldsymbol{\alpha})$ the solution to $\min_{\alpha_0, \boldsymbol{\alpha}} \|\mathbf{y} - (\alpha_0 \mathbf{1} +$

$X\boldsymbol{\alpha})\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\eta^\eta$, where the $N \times (N-1)$ regression matrix is

$$X = \begin{pmatrix} 1 & \dots & & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & & & \vdots \\ & & & 1 \\ 0 & \dots & & 0 \end{pmatrix}.$$

The columns of the Markov random field-equivalent regression matrix $X$ are not $\rho$-standardized. If we were to face this matrix of regressors (not knowing its Markov random field equivalence), we would follow the recommendation of average-centering and $\rho$-standardizing them as it is commonly done. We would then obtain an awkward estimator. Indeed the associated Markov random field estimate would no longer solve (4) but

$$\min_{\boldsymbol{\mu}} \|\mathbf{y} - \boldsymbol{\mu}\|_2^2 + \lambda \sum_{n=1}^{N-1} |\frac{\mu_{n+1} - \mu_n}{d_n}|^\eta, \tag{5}$$

where $d_n = 1/\sqrt{n - n^2/N}$ for $n = 1, \dots, N-1$ are the diagonal elements of the $\rho$-normalizing matrix $D_\rho$. With the weights $1/d_n$ being smallest near the end points and largest in the center, the $\rho$-standardized estimate would be wigglier near the end points and smoother in the center (see Section 4.1). Clearly, $\rho$-standardization is inappropriate.

If instead we $\Sigma$-standardize the columns of $X$, then the good properties of Markov random field smoothing are preserved since $\Sigma = (X_\bullet' X_\bullet)^{-1}$ is a symmetric bi-diagonal matrix with 2 along the main diagonal and $-1$ on the first sub-diagonal, and consequently $D_\Sigma$ is a constant diagonal matrix with $d_n = \sqrt{2}$ for $n = 1, \dots, N-1$. Moreover with $\Sigma$-standardization, the Markov random field weights $(-1/\sqrt{2}, 1/\sqrt{2})$ are the normalized Haar wavelets.

### 3.2 Wavelet smoothing

We now check that Σ-standardization is compatible with two wavelet-based smoothers. *Waveshrink* (Donoho and Johnstone, 1994), a nonparametric wavelet-based smoother, uses a matrix $X$ with average-centered and orthonormal columns by construction. The least squares wavelet coefficients estimate has the property of independent homoscedastic components since $\hat{\boldsymbol{\alpha}}^{\mathrm{LS}} = X'\mathbf{y} \sim \mathrm{N}(\boldsymbol{\alpha}^{\mathrm{true}}, I)$. Donoho and Johnstone (1994) proposed to bias the estimate by thresholding toward zero by means of the hard or soft thresholding function applied componentwise: $\hat{\boldsymbol{\alpha}}_\lambda = \eta_\lambda(\hat{\boldsymbol{\alpha}}^{\mathrm{LS}})$, where

$$\eta_\lambda^{\mathrm{hard}}(\alpha_p) = \alpha_p \cdot 1(|\alpha_p| \geq \lambda) \quad \text{and} \quad \eta_\lambda^{\mathrm{soft}}(\alpha_p) = \mathrm{sign}(\alpha_p)(|\alpha_p| - \lambda)_+. \tag{6}$$

With these methods, any least squares coefficient smaller in magnitude than $\lambda$ is declared not significantly different from zero and set to zero. It is important to note that the hard-estimate corresponds to *best subset variable selection*, and that the soft-estimate (like lasso) solves an $\ell_1$ penalized likelihood problem

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\alpha\|_1, \tag{7}$$

where the isotropic penalty reflects the homoscedasticity of the least squares estimate. In this situation, Σ-standardization retains isotropy since $D_\Sigma$ is the identity.

*Basis pursuit* (Chen, Donoho, and Saunders, 1998) is an extension of *soft-Waveshrink* when the set of wavelets is overcomplete, for instance the union of orthonormal wavelets (e.g., wavelet packets, cosine packets). In that situation the matrix $X$ in (7) has more columns than rows. With a concatenated matrix $X = [X_1 \mid \ldots \mid X_Q]$ of $Q$ orthonormal matrices $X_q$, the least squares estimate is not uniquely defined, but the minimum $\ell_2$-norm $\hat{\alpha}^{\mathrm{LS}} = X'(XX')^{-1}\mathbf{y}$ being uniquely defined, we can calculate the Σ standardization matrix $D_\Sigma^2 = \sigma^2\mathrm{diag}(X'(XX')^{-1}(XX')^{-1}X) = \frac{\sigma^2}{Q^2}I$. We see that Σ-standardization keeps isotropy of the $\ell_1$ prior for *Basis pursuit*.

### 3.3 Principal component regression

We now see how $\Sigma$-standardization leads to a new version of *Principal component regression* (PCR). PCR operates regularization in a different basis than the canonical basis used so far (Massy, 1965). Consequently, the estimated model is interpretable only as a linear combination of the covariates. The new basis $U$ is obtained from the singular value decomposition of $X' = UDV'$, where both $U$ and $V$ are orthonormal and $D$ is a diagonal matrix of singular values organized from largest to smallest. Hence PCR considers the least squares problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - XU\boldsymbol{\beta}\|_2^2, \tag{8}$$

where $XU$ has the advantage of being orthogonal since $(XU)'(XU) = D^2$; here $X$ and $\mathbf{y}$ are assumed average-centered. Based on the independence of the components of $\hat{\boldsymbol{\beta}}^{\mathrm{LS}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2 D^{-2})$, Massy proposed two regularization methods, $\mathrm{PCR}_1$ and $\mathrm{PCR}_2$, the second one being too often forgotten:

- $\mathrm{PCR}_1$: keep the first $k$ coefficients because they have the smallest variance (i.e., largest singular values);

- $\mathrm{PCR}_2$: keep the $k$ most significant components. They are the largest entries in magnitude in $D\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$, since the entries of $D\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$ are homoscedastic and independent. This method is equivalent to *best subset variable selection* in the eigen basis $U$, and therefore has a closed form expression through the hard shrinkage function (6):

  $$\hat{\boldsymbol{\beta}}^{\mathrm{PCR}_2} = D^{-1}\eta_\lambda^{\mathrm{hard}}(D\hat{\boldsymbol{\beta}}^{\mathrm{LS}}).$$

  The rescaling operated with the $D$ matrix allows equitable thresholding between least squares coefficients.

Based on $\Sigma$-standardization and by analogy with *Waveshrink* and *lasso*, we propose a third version of PCR:

- $\mathrm{PCR}_3$: we $\Sigma$-standardize $XU$ with $D_\Sigma = D^{-1}$ and then define the lasso-PCR

as the solution to $\min_{\boldsymbol{\beta}} \|\mathbf{y} - XUD_\Sigma\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$. This new $\Sigma$-standardized lasso-PCR estimate has a closed form expression:

$$\hat{\boldsymbol{\beta}}^{\mathrm{PCR}_3} = D_\Sigma \eta_\lambda^{\mathrm{soft}}(D_\Sigma^{-1}\hat{\boldsymbol{\beta}}^{\mathrm{LS}}).$$

In the second and third versions of PCR, $\Sigma$-standardization shrinks the least squares estimate equitably. Note that when the regression matrix has orthogonal columns, then $\Sigma$-standardization matches $\rho$-standardization.

# 4 Application

## 4.1 Markov random field smoothing

To visually assess the advantage of $\Sigma$-standardization over $\rho$-standardization already demonstrated mathematically in Section 3.1 for Markov random field smoothing, we simulate Gaussian data from the `heavisine` function (Donoho and Johnstone, 1994) represented on top of Figure 1. We then smoothed the data either with $\rho$-standardization by solving (5) or with $\Sigma$-standardization by solving (4). Looking at both smoothed curves represented in the middle and bottom graphs of Figure 1, we observe that the estimate based on $\rho$-standardization becomes wigglier near the end points causing the squared error loss to increase by 30% compared to $\Sigma$-standardization, as expected.

## 4.2 Fuel octane data

Predicting the octane level of a fuel sample from spectrometer measurements is based on a large number of covariates, here $P = 351$, thanks to increasing spectrometer resolution. Data collection is expensive however, so the number of samples $N = 434$ is typically of the order of $P$. The top panel of Figure 2 plots the 434 spectral lines of predictors: we observe that they are very collinear. *Partial least squares* (PLS) is the regression method of choice for chemometricians, and will serve as our benchmark. We also consider four other estimators of the linear model: forward *stepwise* (forward-backward search is computationally too intensive) driven by

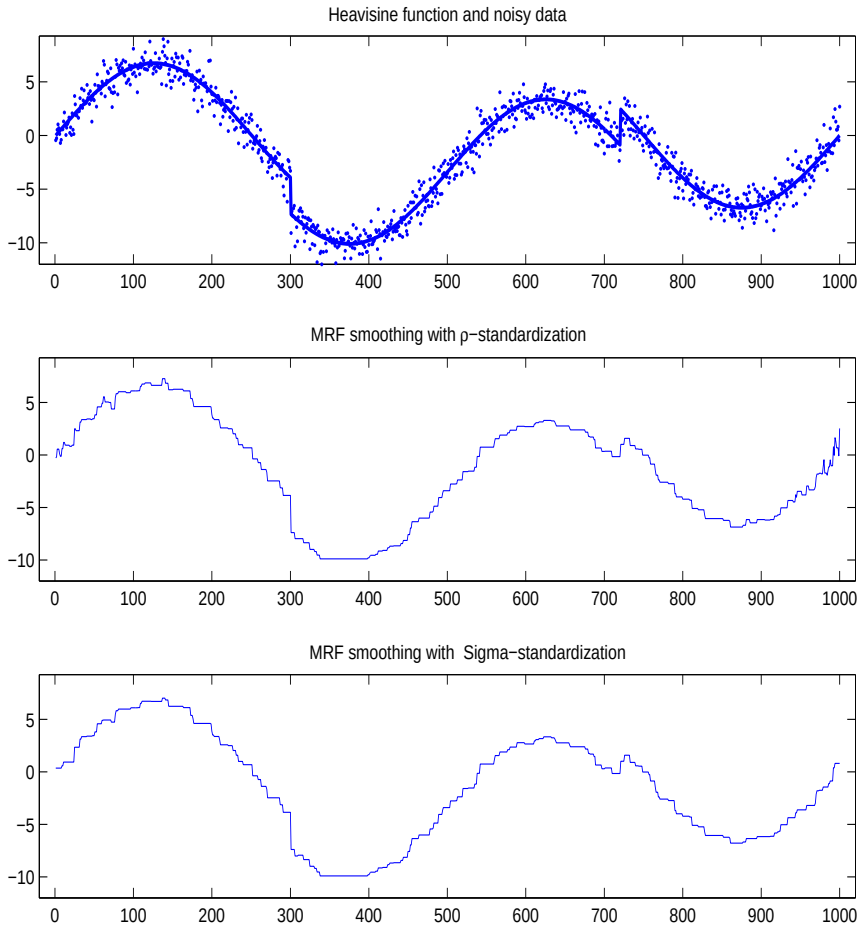


Figure 1: Markov random field smoothing. Top: underlying curve (`heavisine` function) and raw data. Middle: smoothed estimate with $\rho$-standardization. Bottom: smoothed estimate with $\Sigma$-standardization.

BIC (Schwarz, 1978) is the conceptually simplest and most commonly used method, *boosting* driven by AIC is a recent and promising method available in `R` (`mboost` library), $\rho$-standardized *lasso* driven by BIC is the default method available in `R` (`glmpath` library), and $\Sigma$-standardized *lasso* driven by the sparsity $\ell_1$ information criterion ($\mathrm{SL}_1\mathrm{IC}$) is the proposed rescaling for *lasso*.

The estimated coefficients of the four fits are plotted in Figure 2. We observe that the first four estimates are a sparse coefficients sequence and that the last one (partial least squares) is a radically different and smooth sequence. Among

the four sparse estimates, Σ-standardized *lasso* is the least sparse. To compare

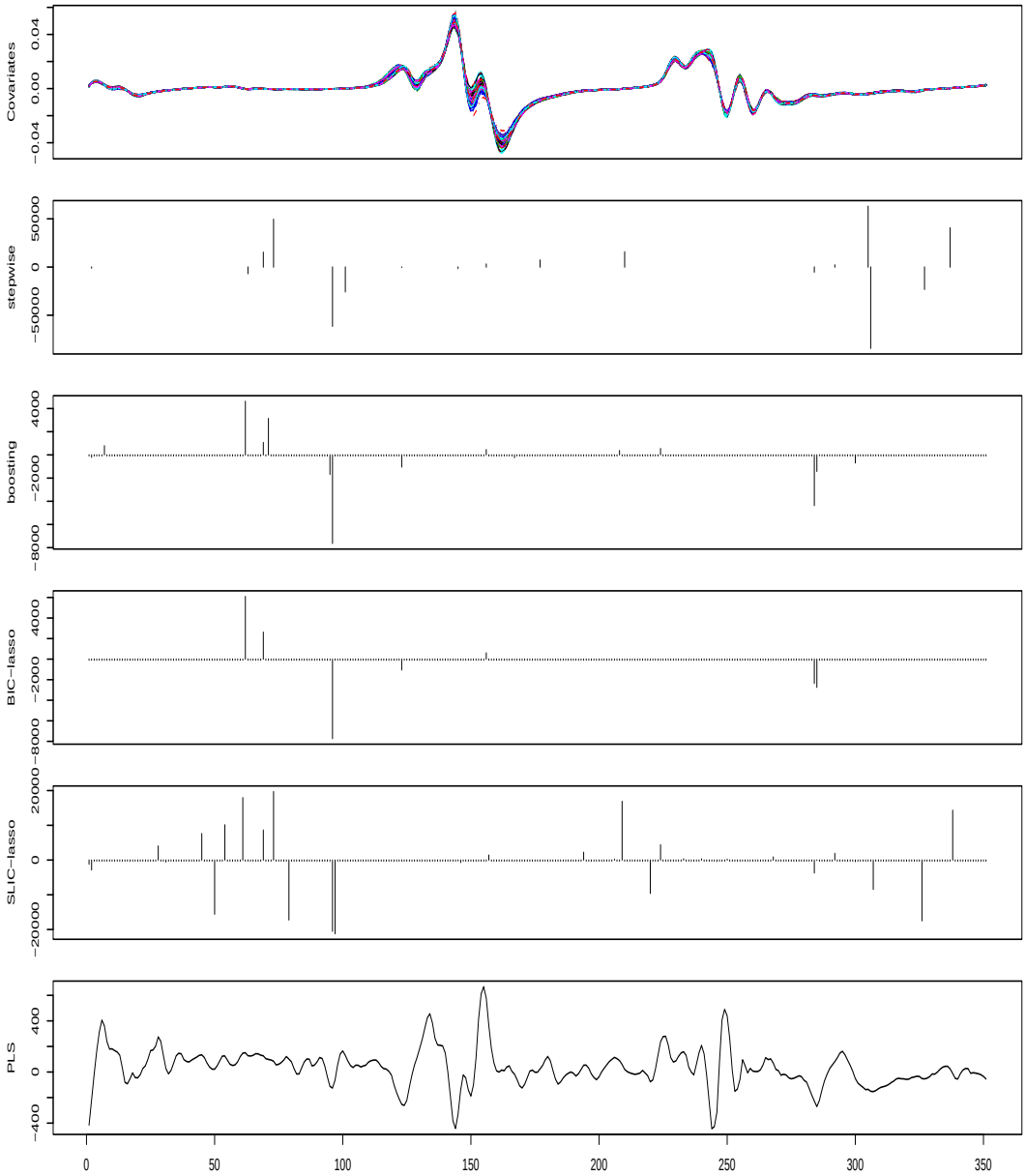


Figure 2: Top graph: $N = 434$ vectors of length $P = 351$ of covariates plotted against their index (i.e., wavelength). Other graphs: estimated coefficients by *forward-stepwise*, *boosting*, $\rho$-standardized *lasso*, Σ-standardized *lasso* and *partial least squares*.

the relative predictive performance of the methods, we perform a Monte Carlo simulation by randomly splitting the data 100 times into two sets of equal size 217, the first one to train, the second to test based on average predictive squared errors. The mean of the 100 predictive measures are reported in the last column

of Table 1, and the corresponding boxplots are shown in Figure 3. Based on this Monte Carlo simulation, *lasso* with $\Sigma$-standardization not only beats *lasso* with $\rho$-standardization, but also beats PLS in terms of predictive performance. Note that we do not report the least squares estimate's performance since it is not uniquely defined with $P \geq N/2$ in the cross-validated training sets.

Table 1: Chemometrics data: Estimated predictive squared error performance (PSE) for five regression methods.

| | stepwise-BIC | boosting | $\rho$-lasso-BIC | $\Sigma$-lasso-$SL_1$IC | PLS |
|---|---|---|---|---|---|
| PSE | 5.28 | 5.00 | 5.04 | 4.72 | 5.10 |

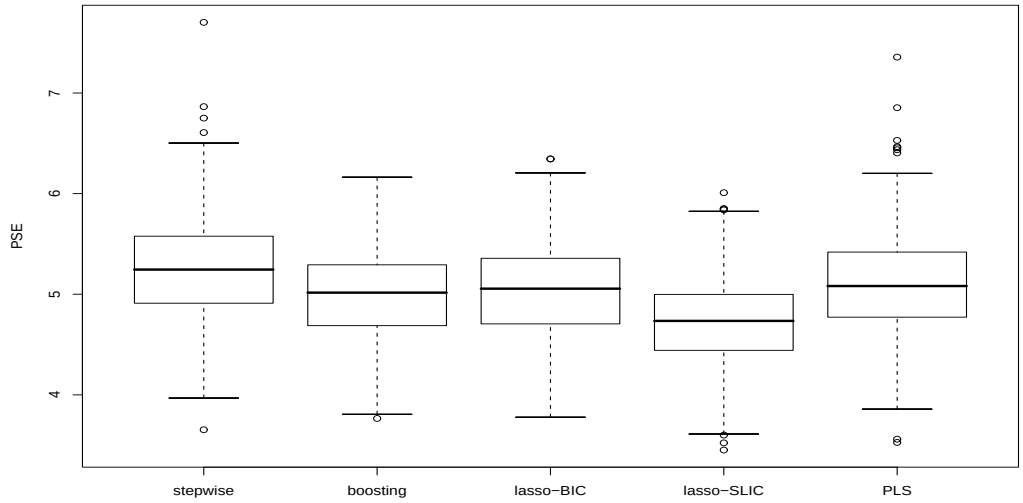


Figure 3: Monte Carlo results for chemometrics data: boxplots of estimated predictive squared errors for five estimation methods.

The information criterion used by *lasso* requires estimation of $\sigma^2$, the variance of the noise measurements. Since the regression matrix has too many columns, the estimate of variance $\hat{\sigma}^2 = \sum_{n=1}^{N}(y_n - \hat{y}_n^{\mathrm{LS}})^2/(N-P)$ is biased downward, and in the extreme situation of more columns than rows, we even have $\hat{\sigma} = 0$. In this situation, we propose a new estimator of variance based on a similar idea of taking the median absolute deviation (MAD) of the first order differences rescaled to be unbiased for Gaussian noise (Donoho and Johnstone, 1995) which is used

in nonparametric regression. For parametric regression, let $X' = UDV'$ be the singular value decomposition used by PCR in Section 3.3. In the $U$ basis, the least squares solution to (8) is $\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = U'X'\mathbf{y} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2 D^{-2})$. Assuming that most true coefficients $\boldsymbol{\beta}$ are small in eigen directions of small spread (small singular values), then most entries in $\hat{\boldsymbol{\gamma}}^{\mathrm{LS}} = D\hat{\boldsymbol{\beta}}^{\mathrm{LS}} \sim \mathrm{N}(\boldsymbol{\gamma}, \sigma^2 I)$ are negligible, like most wavelet coefficients are negligible in wavelet smoothing. So most entries in $\hat{\boldsymbol{\gamma}}^{\mathrm{LS}}$ are essentially noise with variance $\sigma^2$. Hence an estimate of $\sigma$ is the MAD of $\hat{\boldsymbol{\gamma}}^{\mathrm{LS}}$ rescaled for Gaussian noise.

### 4.3 Body fat data

For $N = 71$ healthy German female subjects, body fat measurements and $P = 9$ anthropometric measurements are available for predictive modeling of body fat (Garcia, Wagner, Hothorn, Koebnick, Zunft, and Tippo, 2005). Collecting all 9 measurements is too expensive for practical purposes. To select a subset of covariates while improving bias–variance trade-off, we employ four variable selection methods commonly used in the $P \ll N$ setting: forward-backward *stepwise* driven by BIC, *boosting*, $\rho$-standardized *lasso* driven by BIC, and $\Sigma$-standardized *lasso* driven by $\mathrm{SL}_1\mathrm{IC}$. We use least squares as a benchmark.

The estimated coefficients of the four fits are reported in Table 2. To compare the relative predictive performance of the methods, we perform a Monte Carlo simulation by randomly splitting the data 500 times into two sets of size 35 and 36, the first one to train, the second to test by computing the average predictive squared errors. The mean of the 500 predictive measures are reported in the last column of Table 2, and the corresponding boxplots are shown in Figure 4. Based on this simulation, *stepwise* performs best on this data set with a small number of

Table 2: Body fat data: Estimated coefficients ($\times 100$) of the nine covariates by five methods, and their predictive squared error performance (PSE).

| | $\hat{\alpha}_1$ | $\hat{\alpha}_2$ | $\hat{\alpha}_3$ | $\hat{\alpha}_4$ | $\hat{\alpha}_5$ | $\hat{\alpha}_6$ | $\hat{\alpha}_7$ | $\hat{\alpha}_8$ | $\hat{\alpha}_9$ | PSE |
|---|---|---|---|---|---|---|---|---|---|---|
| stepwise-BIC | 0 | 20.4 | 35.5 | 0 | 181 | 0 | 713 | 0 | 0 | 12.1 |
| boosting | 1.4 | 19.0 | 35.2 | -38.4 | 174 | 332.7 | 366 | 59.5 | 0 | 13.9 |
| $\rho$-lasso-BIC | 0 | 19.5 | 35.0 | 0 | 155 | 153 | 523 | 26.7 | 0 | 14.0 |
| $\Sigma$-lasso-$\mathrm{SL}_1$IC | 1.3 | 19.4 | 35.5 | 0 | 69.2 | 34.5 | 40.5 | 38.2 | 0 | 13.7 |
| least squares | 2.0 | 21.0 | 34.4 | -41.2 | 176 | 574 | 987 | 38.7 | -658 | 15.4 |

covariates. More important is the comparison between rescaling methods employed with the *lasso*. Σ-standardization is better than $\rho$-standardization for lasso, but the improvement in prediction is not as drastic as in the chemometrics example. A pairwise $t$-test between the two methods reveals that Σ-standardization is significantly better; so is Σ-standardization-lasso with *boosting*.

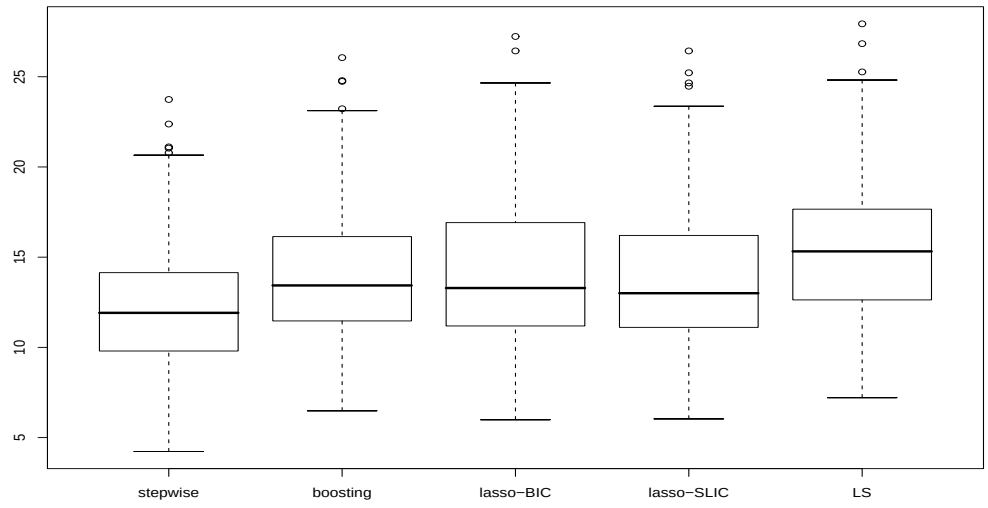


Figure 4: Monte Carlo results for bodyfat data: boxplots of predictive squared errors for five estimation methods.

## 4.4 Poisson smoothing

Additive noise has been assumed so far. We now illustrate how Σ-standardization operates a proper rescaling for heteroscedastic Poisson data: $\mathbf{s} \sim \text{Poi}(\boldsymbol{\mu})$. Many smoothers have been developed to smooth Poisson signals using wavelets (see Besbeas, De Feis, and Sapatinas (2004) for a review). Wavelet-based nonparametric linear models for Gaussian data assume that the underlying signal $\boldsymbol{\mu} = X_0\boldsymbol{\alpha}_0 + X\boldsymbol{\alpha}$, where $X_0$ and $X = X_\bullet$ are the matrices of discretized ortho*normal* father and mother wavelets, respectively (Donoho and Johnstone, 1994). To account for the heteroscedasticity induced by Poisson noise, many wavelet-based estimators for Poisson data rescale the wavelet matrix (Kolaczyk, 1999; Sardy, Antoniadis, and

Tseng, 2004; Fryzlewicz and Nason, 2004).

An alternative approach is to use any wavelet matrix, regardless whether normalized or not, and employ $\Sigma$-standardization to account for heteroschedasticity in the following way. Since $\mathbf{y} \sim \text{Poi}(\boldsymbol{\mu})$, then $\hat{\boldsymbol{\alpha}}^{\text{MLE}} = X'\mathbf{y}$ is an unbiased estimate of the wavelet coefficients with covariance matrix $\Sigma = X'\text{diag}(\boldsymbol{\mu})X$. We therefore propose to $\Sigma$-standardize and use the rescaled wavelet matrix $XD_\Sigma$, where $D_\Sigma^2 = \text{diag}(\Sigma)$. In practice, $D_\Sigma$ must be estimated iteratively starting with $\hat{D}_\Sigma^2 = \text{diag}(X'\text{diag}(\mathbf{y})X)$, and updating with $\hat{D}_\Sigma^2 = \text{diag}(X'\text{diag}(\hat{\boldsymbol{\mu}})X)$. The rescaling is therefore adaptive to the underlying signal.

We illustrate $\Sigma$-standardized Poisson smoothing with data measured by the burst and transient source experiment (BATSE) instruments on board of NASA's Compton Gamma Ray Observatory. The device measures arrival times of high energy gamma rays. Using a partition of time, the data consists of counts of gamma rays in each bin; for details see Meegan, Fishman, Wilson, Paciesas, Pendleton, Horack, Brock, and Kouveliotou (1992). We focus in particular on the trigger 551 data used by Besbeas, De Feis, and Sapatinas (2004) to illustrate how different estimators can adapt to the sharp features of the underlying signal. The signal recorded during 0.94 seconds has length $N = 1024$, but as pulses occurred in the first half (until 0.47 seconds), we show only this half on Figure 5 to zoom on the relevant area.

Figure 5 shows the raw data along with three estimates: translation invariant Bayesian MultiScale Model (BMSM – TI) (Kolaczyk, 1999) based on Haar wavelets is the first wavelet-based estimator employed on these data, translation invariant $\ell_1$ penalized likelihood (Sardy, Antoniadis, and Tseng, 2004) using $\Sigma$-standardized Haar wavelets is the proposed rescaling method (other wavelets could be used with $\Sigma$-standardization), and Markov random field $\ell_1$ smoothing is a competitor to wavelet-based methods (estimator of the form (4) with Poisson likelihood (Sardy and Tseng, 2004)). All three methods are available in `Matlab`, and are designed to detect local features: in particular the two dips at times 0.22 and 0.24 are preserved. We observe that two estimators, using $\Sigma$-standardized wavelets and

Markov random field, hint for potential peaks not revealed by BMSM that seems to oversmooth in regions of low Poisson intensities. Σ-standardization is a simple and efficient way of taking heteroscedasticity into account.

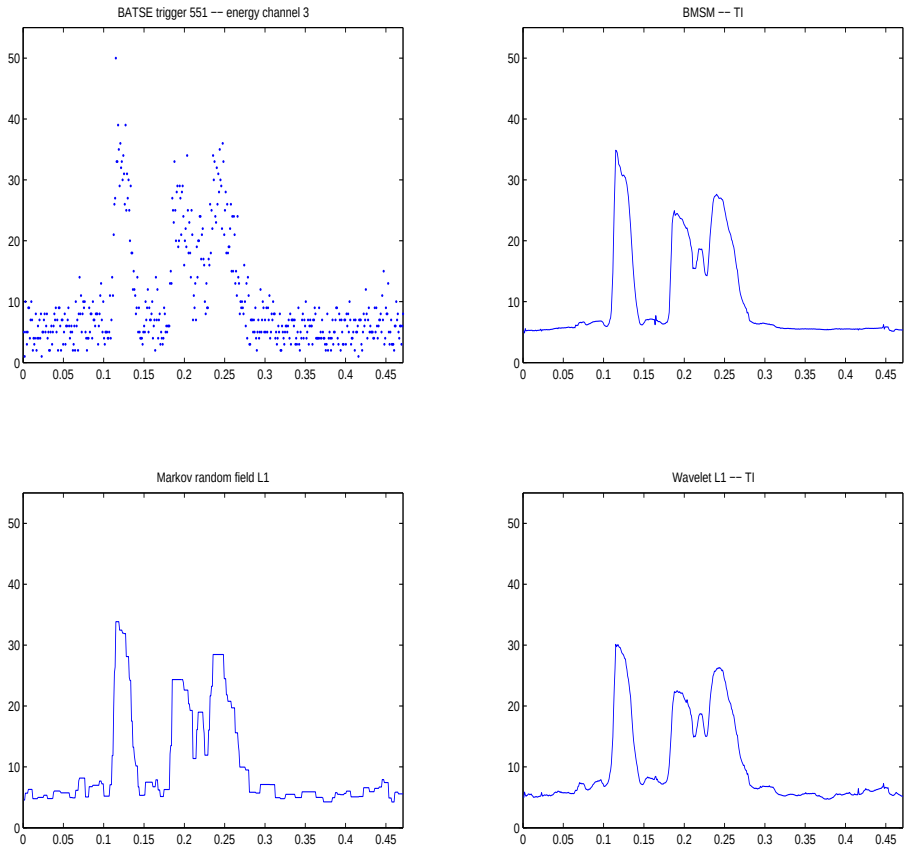


Figure 5: Poisson wavelet smoothing. Top left: BATSE trigger 551 data until 0.47 seconds. Top right: BMSM using orthogonal wavelets. Bottom right: $\ell_1$ penalized likelihood using Σ-standardization. Bottom left: Markov random field $\ell_1$. (TI stands for translation invariant)

## 5 Conclusion

The new method of rescaling covariates is based on the variance of the MLE rather than the variance of the covariates. Based on a heuristic, illustrative examples, and applications to real data sets, we claim that Σ-standardization should be employed

before using an isotropic prior (e.g., lasso, ridge regression, or Markov random field), and should be considered as an alternative to the commonly used $\rho$-standardization whether or not the covariates are in the same units, and whether employing a parametric or a nonparametric estimator.

## 6 Acknowledgement

I would like to thank two referees for their valuable comments, and Arne Kovac for a stay at the University of Bristol during which part of this work was developed.

## References

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B* **48**, 192–236.

Besbeas, P., De Feis, I., and Sapatinas, T. (2004). A Comparative Simulation Study of Wavelet Shrinkage Estimators for Poisson Counts. *International Statistical Review* **72**, 209–237.

Buehlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*(4).

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20**, 33–61.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Friedman, J. H. (1991). Multivariate adaptive regression splines (Disc: P67-141). *The Annals of Statistics* **19**, 1–67.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (C/R: P23-39). *Technometrics* **31**, 3–21.

Fryzlewicz, P. and Nason, G. (2004). A Haar-Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics* **13**, 621–638.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Garcia, A. L., Wagner, K., Hothorn, T., Koebnick, C., Zunft, H.-J. F., and Tippo, U. (2005). Improved prediction of body fat by measuring skinfold thickness circumferences, and bone breadth. *Obesity Research* **13**, 626–634.

Geman, S. and Geman, D. (1984). Stochastic relaxation. Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **61**, 721–741.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Jensen, D. R. and Ramirez, D. E. (2008). Anomalies in the Foundations of Ridge Regression. *International Statistical Review* **76**, 89–105.

Kolaczyk, E. (1999). Bayesian multi-scale models for Poisson processes. *Journal of the American Statistical Association* **94**, 920–933.

Marquardt, D. W. (1980). Comments on "A critique of some ridge regression methods". *Journal of the American Statistical Association* **75**, 87–91.

Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice. *The American Statistician* **29**, 3–20.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association* **60**, 234–256.

Meegan, C. A., Fishman, G. J., Wilson, R. B., Paciesas, W. S., Pendleton, G. N., Horack, J. M., Brock, M. N., and Kouveliotou, C. (1992). The Spatial Distribution of Gamma Ray Bursts Observed by BATSE. *Nature* **355**, 143–145.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models.

*Journal of the Royal Statistical Society, Series A* **135**, 370–384.

Sardy, S., Antoniadis, A., and Tseng, P. (2004). Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics* **13**, 399–421.

Sardy, S. and Tseng, P. (2004). On the statistical analysis of smoothing by maximizing dirty markov random field posterior distributions. *Journal of the American Statistical Association* **99**, 191–204.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association* **75**, 74–81.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological* **58**, 267–288.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59. Series in Applied Mathematics, SIAM, Philadelphia.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiah (Ed.), *Multivariate Analysis*, pp. 391–420. Academic Press.