Reno: regularized non-parametric analysis of protein lysate array data

Bin Li¹, Feng Liang¹, Jianhua Hu²,*, and Xuming He³

¹Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, 61820, USA, ²Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston, TX, 77230, USA and ³Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, USA

ABSTRACT

Motivation: The reverse-phase protein lysate arrays have been used to quantify the relative expression levels of a protein in a number of cellular samples simultaneously. To avoid quantification bias due to mis-specification of commonly used parametric models, a non-parametric approach based on monotone response curves may be used. The existing methods, however, aggregate the protein concentration levels of replicates of each sample, and therefore fail to account for within-sample variability.

Results: We propose a method of regularization on protein concentration estimation at the level of individual dilution series to account for within-sample or within-group variability. We use an efficient algorithm to optimize an approximate objective function, with a data-adaptive approach to choose the level of shrinkage. Simulation results show that the proposed method quantifies protein concentration levels well. We show through the analysis of protein lysate array data from cell lines of different cancer groups that accounting for within-sample variability leads to better statistical analysis.

Availability: Code written in statistical programming language R is available at: http://iddn.mdacc.tmc.edu/∼jhu/Reno

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on June 23, 2011; revised on March 10, 2012; accepted on March 12, 2012

1 INTRODUCTION

Protein microarray technologies [e.g. Cahill and Northoll (2003); Ivanov et al (2008); MacBeath and Schreiber (2000)] have been developed to measure protein concentrations in a high-throughput fashion. Extensive reviews of this technology can be found in Borrebæk and Wingren (2007) and Prent et al (2008). A single nitrocellulose-coated array slide can measure concentrations of a common protein in hundreds of samples in the form of dilution series. The samples are hybridized and label-attached with primary and biotinylated secondary antibodies and the protein concentrations are then measured using streptavidin-linked labels that bind to the biotin. The final product of each array is an image file in which quantified spots represent the observed protein expression levels at various dilution steps.

Protein lysate array technology has shown its promise in a number of clinical studies [e.g. Piniset al (2008); Sahin et al (2007)]. In particular, its applications to various cancer studies have been documented extensively; see, for example, Cat et al (2013); Sager et al (2008); Cheng et al (2008); Stoeve et al (2008); Kimm et al (2008); Sourixture et al (2008); Shibes et al (2008). Various procedures to improve the analysis of protein lysate arrays have been proposed recently. For example, Brese et al (2010) proposed antibody-mediated signal amplification to increase the sensitivity of this technology. Neeley et al (2009) introduced a variable slope normalization among arrays to help reduce loading bias and recover true correlation structures among proteins.

In this article, we focus on the protein quantification problem of estimating the relative protein concentrations in the arrayed samples. In this area, the commercial analysis package MicroVigene (http://www.vigenetech.com/products.htm) estimates the protein expression level by fitting a four-parameter logistic model to each dilution series. Kreutz et al (2007) used a linear model between the observed protein expressions (or at logarithmic scale) and the underlying concentration levels. Tabus et al (2004) modeled the mean of the observed expression level as a sigmoidal curve and estimated the model parameters via the non-linear least squares. Alternatively, Zhang et al (2004) modeled the serial dilution curve based on the Sips model (similar to the logistic model) to characterize the relationship between signals in successive dilution steps. These methods can be considered as parametric.

On the contrary, Hu et al (2007) proposed a non-parametric approach by assuming that the median of the observed protein expression is equal to a monotonically increasing function without a parametric form. This non-parametric method, contained in a recently developed statistical tool RPPAnalyzer for analyzing protein lysate array data (Manneper et al (2010), is highly data-adaptive without bias due to mis-specification of \( g \). Simulation studies and real experiments have shown the advantage of the non-parametric approach in producing robust results in a variety of scenarios. However, the existing methods of estimation deteriorate in performance as the noise to signal ratio increases in the data. In addition, the non-parametric estimates, if used to estimate protein concentration at the level of individual dilution series, tend to be unstable due to data sparsity.

Our work aims at improving the accuracy of protein level quantification over the non-parametric procedure described in Hu et al (2007) by introducing a data-adaptive approach to choose the level of shrinkage.
is the corresponding dilution index at the \( t \)-th step. The relationship between \( y_{ij} \) and the unobserved protein concentration level \( x_{ij} \) (at the log2 scale) can be modeled as

\[
y_{ij} = g(x_{ij} - d) + \epsilon_{ij},
\]

where \( g \) is the protein-specific response curve, \( d_i \) defines the corresponding dilution index at the \( i \)-th step and \( \epsilon_{ij} \) denotes random noise. In a typical dilution series, \( d_i - d_{i-1} = 1 \). In this formulation, we allow each replicate to have its own protein concentration level \( x_{ij} \), which differs from the set-up of [Hu et al. 2009], where \( x_{ij} = x_0 \) for all the replicates of the \( i \)-th sample. If we adopt the method of [Hu et al. 2009], we can estimate a non-parametric estimate of \( g \), as well as the concentration levels \( x_{ij} \), by minimizing

\[
\sum_{i,j,l} |y_{ijl} - g(x_{ijl} - d_l)| + \lambda \max_x |g'(x)|,
\]

for some smoothing parameter \( \lambda \), subject to the constraint that \( g \) is a non-decreasing function. Note that \( x_{ij} \) are identifiable only up to a constant, and the relative differences between \( x_{ij} \)'s are of interest in protein concentration.

As pointed out by Yang and He 2011, the complexity of this quantification problem comes from the dimension of \( x_{ij} \), which increases with the sample size. One assumption made in earlier work is that \( x_{ij} \) are constant within the \( i \)-th sample. This assumption is practically useful because it reduces the dimensionality of the problem and ensures stable estimates. However, it could lead to bias in estimation, and more importantly, could mask or distort variability across replicates.

We propose to allow \( x_{ij} \) to vary with \( j \) but regularize the estimation problem by using a penalty term on within-sample variabilities. As a result, we shrink some of the replicate-level estimates to common values but allow within-sample variabilities to be retained in other samples as needed. As with other regularization methods in statistics, our proposed method aims to balance the bias-variance trade-off in a data-adaptive way.

The rest of the article is organized as follows. We describe the new method, Regularized Nonparametric (Reno) analysis of lysate arrays, in Section 2. We demonstrate the performance of the new estimator through simulation studies in Section 3, and show how the proposed method leads to better significance testing in a real data example in Section 4. Some concluding remarks are given in Section 5.

2 METHODS

2.1 New objective function

We formulate the problem of simultaneously estimating \( g \) and the vector \( \mathbf{x} = (x_i : i = 1, \ldots, m, j = 1, \ldots, n_i) \) as follows

\[
\min_{g, \mathbf{x}} \sum_{i,j,l} L_i(y_{ijl}, g(x_{ijl} - d_l)) + \lambda_1 V_1(g) + \lambda_2 V_2(\mathbf{x})
\]

where

\[
L_i(y_{ijl}, g(x_{ijl} - d_l)) = |y_{ijl} - g(x_{ijl} - d_l)| + \lambda_1 V_1(g) + \lambda_2 V_2(\mathbf{x})
\]

is the \( L_i \) loss function measuring the discrepancy between the observed expression level \( y_{ijl} \) and its fitted value, and \( V_1 \) and \( V_2 \) are penalty functions on \( g \) and \( \mathbf{x} \), respectively, with \( \lambda_1 \) and \( \lambda_2 \) to be determined. As in [Hu et al. 2009], we choose the penalty on \( g \)

\[
V_1(g) = \max_x |g'(x)|,
\]

which leads to a quadratic spline solution of \( g \) [Koenker, 2005]. The \( L_i \) loss function is appealing for lysate data quantification, because the unknown function \( g(x) \) corresponds to the median protein expression level, and the relationship is equivariant under any monotone transformation of the response.

The new penalty term \( V_2(\mathbf{x}) \) used in (3) aims to regularize the solution of a high-dimensional vector \( \mathbf{x} \). Empirical results show that the optimization problem without this penalty (that is, \( \lambda = 0 \)) is often ill-conditioned and leads to unstable results. The special case of \( \lambda_2 = \infty \) corresponds to the assumption in earlier work that \( x_{ij} = x_0 \), for all \( j = 1, \ldots, n_i \). The results obtained under this setting mask the variability across replicates.

In protein microarray experiments, it is often the case that the samples can be divided into \( G \) subgroups \( G_1 \cup \cdots \cup G_G \), and the protein concentration levels are similar within each subgroup. In this article, we focus on a common case where each subgroup contains biological or technical replicates of some kind. In our notation, all the replicates within the sample form a subgroup. Naturally, we use the following penalty on \( \mathbf{x} \),

\[
V_2(\mathbf{x}) = \sum_{j=1}^{m} \sum_{l=1}^{G_j} |x_{ijl} - x_{j'}| = |\mathbf{A}\mathbf{x}|,
\]

Simultaneous optimization over \( g \) and \( \mathbf{x} \) is numerically challenging, and we propose an approximation to (3) in the next subsection.

2.2 Modified objective function

A natural way to solve (3) is a block-wise coordinate descent algorithm that iterates between the following two steps:

- **Step 1:** minimize (3) over \( g \) given \( \mathbf{x} \);
- **Step 2:** minimize (3) over \( \mathbf{x} \) given \( g \).

Iterative algorithms of this type have been widely adopted [Hu et al. 2009] and [Hubisz et al. 2010].

Step 1 reduces to

\[
\min_g \sum_{i,j,l} L_i(y_{ijl}, g(x_{ijl} - d_l)) + \lambda_1 \max_x |g'(x)|
\]

subject to the constraint that \( g \) is a non-decreasing function. This is the same problem considered by [Hu et al. 2009] and [Hubisz et al. 2010], and we use the same procedure as before. The solution of \( g \) is a quadratic spline, and it can be found via linear programming with the R package Cobs developed by [Hu and He, 2009].

Given \( g \), Step 2 solves

\[
\min_{\mathbf{x}} \sum_{i,j,l} L_i(y_{ijl}, g(x_{ijl} - d_l)) + \lambda_2 |\mathbf{A}\mathbf{x}|,
\]

which can be optimized for each \( j \) separately. Still, the non-linear nature of \( g \) makes this problem harder. We propose to simplify the calculation by using an approximate loss function. Note that

\[
|y_{ijl} - g(x_{ijl} - d_l)| = |g(g^{-1}(y_{ijl} - d_l)) - g(g^{-1}(x_{ijl} - d_l))| = |x_{ijl} - g^{-1}(y_{ijl}) - g^{-1}(x_{ijl})|.
\]

Then for the \( i \)-th subgroup, we solve, in lieu of (3),

\[
\min_{\mathbf{x}} \sum_{j=1}^{m} \sum_{l=1}^{G_j} \lambda_2 |\mathbf{A}\mathbf{x}|
\]

where

\[
\lambda_2 = |g(g^{-1}(y_{ijl} - d_l)) - g(g^{-1}(x_{ijl} - d_l))| = |x_{ijl} - g^{-1}(y_{ijl}) - g^{-1}(x_{ijl})|
\]

are calculated based on the current values of \( g \) and \( x_{ijl} \)'s from the previous iteration. What makes this approximation attractive is that the optimization (3) can be solved by weighted median quantile regression using the R package quantreg [Koenker, 2005].

In summary, our proposed method uses the following steps:
The parameter $\lambda_2$ decreases of the objective function. We also note that centering indeed decreases the original objective function. If not, we do a grid search where validation data for that case. To accommodate the special structure here, we perform simulation studies to compare our procedure \cite{Hu et al. 2008}.

In each iteration, we add a step to check whether the new estimates indeed decrease the original objective function. If not, we do a grid search where validation data for that case. To accommodate the special structure here, we perform simulation studies to compare our procedure \cite{Hu et al. 2008}.

Throughout this section, we evaluate the estimation error by

$$Err(\hat{x}, x) = \sum_{l=1}^{g} \sum_{d=1}^{t} \sum_{g'} \left| \gamma_{g'}(x_{ijl}) - \gamma_{g'}(\hat{x}_{ijl}) \right|,$$

where $\gamma(x)$ is the true response function. Note that $|\gamma(x) - \gamma(z)| \approx \| \gamma'(x) \| |x - z|$, so this criterion downweights the estimation error of $x$ where $|\gamma'(x)|$ is small, i.e. in the nearly flat regions of $x$. Obviously, accuracies in the nearly flat regions of $g$ cannot be expected from any method. We choose Oracle as the benchmark method and report the relative error, the ratio of the error of each method to that of Oracle.

### 3.1 Experiment 1

In our first experiment, the response curve is taken from two sigmoid curves in the form of $g(x) = \alpha + \beta/(1 + 2^{-x})$ with $\alpha = 3000$ and $\beta = 10000$, but $\gamma = 0.7$ for positive $x$ and $\gamma = 2.1$ for negative $x$. For positive $x$, the noise is generated from a scaled $t$ distribution with three degrees of freedom with the scale $\sigma = 6000 \times (x + 5)^{-1}$, for negative $x$, the noise follows the normal distribution $N(0, 600^2)$.

We conducted 100 simulation trials, and the results are summarized in Figure \[\text{Figure 1}\]. For the first 20 datasets, the tuning parameter $\lambda_2$ was chosen adaptively by CV. The upper left panel of Figure \[\text{Figure 1}\] is the boxplot of CV($\lambda_2$) based on these 20 datasets. The value $\lambda_2^* = 1500$ is a good choice for most datasets from this model. To save time in the simulation study, we used this fixed value for all other datasets. But we expect the results for Reno to be similar if the CV is done for each trial. The upper right panel of Figure \[\text{Figure 1}\] is the boxplot of the relative errors of each competing method over 100 trials. It is clear that the proposed method Reno outperforms its competitors.

We repeated the data generating process with two smaller error variances, where the SD is reduced by a factor of 2/3 and 1/2, respectively. The results are summarized in the lower half of Figure \[\text{Figure 1}\]. In all these cases, Tabus performed the worst due to model misspecification. The performances of Nonpa and Same are similar. When the SNR (signal-to-noise ratio) is low, even if many of the three replicates have different $x_i$ values, the variance deduction from pooling information from the replicates to estimate a single protein concentration level more than offsets the bias. Thus, Nonpa and Same have smaller errors than Indep in the upper right panel of Figure \[\text{Figure 1}\]. When the SNR is higher, bias becomes dominating, so Indep has a smaller error than Nonpa and Same in the lower two panels of Figure \[\text{Figure 1}\]. It is worth noting that Reno consistently outperforms, and it even beats Oracle when the SNR is low. If the calibration curve $g$ were known, Oracle would have the best performance. However, when $g$ is unknown and the SNR is low, adaptively borrowing information across replicates can improve estimation accuracy of $g$ and hence the quantification accuracy of $x$.

### 3.2 Experiment 2

In the second experiment, we generated the response curve from a single sigmoidal curve with $\alpha = 10000$, $\beta = 4000$ and $\gamma = 0.8$. The errors were generated independently from normal distribution with the following alternatives:

- **Indep**: the procedure with $\lambda_2 = 0$, which is equivalent to estimating a single concentration level for each dilution series.
- **Same**: the procedure with $\lambda_2 = \infty$, which is equivalent to assuming that the replicates within each sample have exactly the same protein concentration levels.
- **Nonpa**: the procedure with $\lambda_2 = \infty$, but $\lambda_0$ is solved using the procedure of \cite{Hu et al. 2008}, instead of the approximation \cite{Tabus et al. 2008}. We include this alternative procedure to evaluate the accuracy of our approximation. It turns out that Same is much faster than Nonpa without much sacrifice in accuracy.
- **Tabus**: the parametric procedure based on sigmoid curve of \cite{Tabus et al. 2008}.
- **Oracle**: the procedure with $\lambda_2 = \infty$ only for the samples with $x_0 = x_1 = \ldots = x_n = 0$.

We include this alternative procedure to evaluate the accuracy of our approximation. It turns out that Same is much faster than Nonpa without much sacrifice in accuracy.
Fig. 1. Results from Experiment 1. The upper left panel displays the boxplot of the CV measures of the first 20 datasets at several values of \( \lambda \). Boxplots for the estimation errors (relative to Oracle) on 100 datasets are displayed in the upper right panel (original data generating process, \( \lambda = 1500 \)), the lower left panel (with the error SD reduced by a factor of \( 2/3, \lambda = 1000 \)), and the lower right panel (with the error SD reduced by \( 1/2, \lambda = 750 \)).

N(0, 5002). The true \( x \) values and the rest of the set-up are the same as in Experiment 1. The result are now presented in Figure 2.

In this experiment, Indep is always inferior to others, confirming that estimating individual \( x_{ij} \) for all replicates is usually not recommended. It is interesting to note that Same and Nonpa outperform Oracle when the SNR is small (the upper right panel of Fig. 2), but not in the other cases. Since the data are generated from a sigmoidal curve, the results of Tabus are quite decent, but Reno remains a top performer.

4 ANALYSIS WITH REAL DATA

We analyzed the protein lysate data from Mendes et al. (2007) with the proposed Reno, as well as the non-parametric procedure Nonpa from Hu et al. (2007) and the parametric procedure Tabus from Tabus et al. (2006). The data consist of the intensity measurements for 90 samples from 52 protein arrays. Each sample has three replicates that are diluted 2-fold six times. The fitted curves from Reno and Nonpa are very similar for all arrays, but the parametric curves look different on some datasets. This is because the curve estimated by Tabus is constrained to take the logistic shape, which may not fit the data well. Between Reno and Nonpa, we noticed that the estimated protein concentration levels \( x_{ij} \) differ on a number of arrays. Those differences have an impact on subsequent statistical analysis, as shown in the example below.

The samples in this study are cell lines from 12 different cancer groups. We use the data to find out whether protein concentration levels are significantly different among cancer groups. We conducted
Fig. 2. Results from Experiment 2. The upper left panel displays the boxplot of the CV measures of the first 20 datasets at several values of $\lambda$. Boxplots for the estimation errors (relative to Oracle) on 100 datasets are displayed in the upper right panel (original data generating process, $\lambda = 800$), the lower left panel (with the error SD reduced by a factor of 2/3, $\lambda = 500$), and the lower right panel (with the error SD reduced by 1/2, $\lambda = 400$).

a two-sample $t$-test based on the average concentration levels $z_i = (x_{i1} + x_{i2} + x_{i3})/3$ by assuming that $z_i$ are normally distributed with mean $\mu$ and variance $\sigma^2 + \tau^2_i$, where $\tau^2_i$ is estimated by the sample variance of the replicates $x_{i1}, x_{i2}$ and $x_{i3}$, and $\sigma^2$ denotes the variance component that is homogenous for all samples from the same cancer group.

For Nonpa, $\tau^2_i = 0$, but for Reno, $\tau^2_i$ can be non-zero, and the MLE $\hat{\mu}$ is a weighted average of $z_i$'s. Samples with higher within-sample variation are down-weighted, because their estimated $x_{ij}$ values are less reliable. As will be shown in the examples below, the within-sample variation, which is available only with the Reno procedure, indeed provides valuable information in the statistical analysis.

We focus on protein cdk7, for which the $P$-values for contrasting the lung cancer group with two other cancer groups (Colon or Sarcoma) are listed in Table 1. The estimates from Nonpa have a hard time detecting the difference between the lung cancer group and the other two groups. To see what led to the difference, we display

<table>
<thead>
<tr>
<th>Protein</th>
<th>Tabus</th>
<th>Nonpa</th>
<th>Reno</th>
</tr>
</thead>
<tbody>
<tr>
<td>cdk7</td>
<td>0.060</td>
<td>0.101</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>0.018</td>
<td>0.042</td>
<td>0.006</td>
</tr>
<tr>
<td>p19</td>
<td>0.048</td>
<td>0.041</td>
<td>0.104</td>
</tr>
<tr>
<td>Zap70</td>
<td>0.020</td>
<td>0.030</td>
<td>0.188</td>
</tr>
</tbody>
</table>

The lung cancer group has 13 cell lines/samples, Colon has 8, Sarcoma has 14, Pancreatic has 8 and Ovarian has 5.
the data of four samples from the lung cancer group in Figure 3(left panel). Sample 1 has higher expressions than the other samples, and more importantly, there is a visible within-sample variation for Sample 1, but not for the other samples. This information is reflected in the estimates of protein concentrations from Reno: 4.2, 1.9 and 1.7 for the three replicates. As a consequence, Sample 1 gets down-weighted in the statistical analysis with the following effects: (i) it decreases the average of the lung cancer group, and therefore enlarges the mean difference between groups; and (ii) it reduces the group variance, making the t-test more powerful. This explains why the P-values from Reno are less than a third of the P-values from Nonpa.

Note that Reno provides more accurate and robust uncertainty analysis than its competitors, but does not always provide smaller P-values. In some cases, Reno points to the opposite direction when Tabus and Nonpa report smaller P-values by ignoring within-sample variations. Results from two other example slides (proteins 1st column), the proposed method is helpful in providing additional information in the subsequent statistical analysis of the lysate array data. Note that the estimated calibration curves from different methods are essentially the same, and the differences are in the concentration estimates of individual samples.

Funding: National Institute of Health Grants R21CA129671, R01GM080503, R01CA158113 and NCI CA97007, and the National Science Foundation Awards DMS-0706818, DMS-0800631 and DMS-1007396.

Conflict of Interest: none declared.

REFERENCES


Regenerated Non-parametric Analysis of Protein Lysate Arrays