Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data

John S. Gustafsson*† Robert Svensson‡ Chris A. Glasbey§
Anders Blomberg‡ Mats Rudemo †

December 9, 2003

Abstract

Two-dimensional gel electrophoresis is a major technique in global analysis at the protein level. This paper examines spot volume data from three gel sets with radioactively labeled yeast *Saccharomyces cerevisiae* proteins. A strong variance versus mean dependence in data is found to be stabilized by applying a shifted logarithmic transformation. However, transformed data show a remaining substantial variance heterogeneity for different proteins. Furthermore, examination of Studentized residuals reveals that transformed data are approximately normally distributed and that there are spatial correlations among the measurement errors in the gel.

Keywords: Reproducibility; Statistical analysis; Transformation; Two-dimensional gel electrophoresis; Variation

1 Introduction

Two-dimensional (2-D) gel electrophoresis (O'Farrell, 1975) is still the most frequently applied method to simultaneously separate and quantitate thousands of proteins from cell and tissue samples (Dowsey et al., 2003). The proteins are separated according to their isoelectric point in a first-dimensional separation (x axis) and their size in a second-dimensional separation (y axis). Gels with spot patterns of labeled or stained proteins are converted to digital images using scanning devices. The gel images are then processed to detect and quantitate the protein spots and to match spots in different gels that correspond to the same protein.

*Correspondence: John S. Gustafsson, Department of Mathematical Statistics, Chalmers University of Technology, SE-412 96 Göteborg, Sweden. **E-mail:** johng@math.chalmers.se **Fax:** +46-31-7723508
†Department of Mathematical Statistics, Chalmers University of Technology
‡Department of Cell and Molecular Biology - Microbiology, Göteborg University
§Biomathematics and Statistics Scotland, Edinburgh, UK

The objective of a statistical analysis of spot volume data can for example be to identify differentially expressed proteins or to perform a cluster or discriminant analysis of proteins or samples. Statistical methods are often based on that data are normally distributed with constant variance not dependent on the mean. It is important to assess that data do not severely violate these assumptions; if so, we have either to transform data or to develop methods for the particular type of data.

In this paper we consider two issues in the analysis of quantitative 2-D gel data: (i) transformations that stabilize the variance versus mean dependence in spot volume data, and (ii) statistical exploration and characterization of transformed data. These considerations both point at possible systematic errors in the preceding image analysis step and guide the choice of methods in a subsequent statistical analysis. We examine data from three gel sets with radioactively labeled yeast *Saccharomyces cerevisiae* proteins. Several studies of the variation in quantitative 2-D gel data exist (Garrels, 1989; Burstin et al., 1993; Blomberg et al., 1995; Voss and Haberl, 2000; Mahon and Dupree, 2001; Molloy et al., 2003). Our study is focused on statistical aspects of the variation.

Quantitative 2-D gel data typically comprise many variables (proteins) but few observations (experimental replications). This is also the case in another large-scale measurement technique in molecular biology: the genome-wide measurement of mRNA levels with microarrays. There has been an extensive work in the statistical community on the design and analysis of gene expression experiments using microarrays (see, for example, Parmigiani et al. (2003); Speed (2003)). An example of a statistical issue that has been considered in the analysis of microarrays, but to our knowledge not yet in the analysis of 2-D gels, is the multiple comparison issue in the hypothesis testing for differential expression (Dudoit et al., 2003). Some aspects in this paper are inspired by the analysis of microarrays and one of our aims is to draw attention to the similarities in the statistical challenges in these two types of genome-wide analyses.

2 Material and methods

2.1 Data sets

The three data sets that are examined in the paper consist of spot volume data from 2-D gels with radioactively labeled proteomes of yeast *Saccharomyces cerevisiae*: (i) a 2×5 gel set that has been used in a genome-wide analysis of the steady-state salt stress response in yeast (Alipour et al., 2003) (ii) a $2\times2\times3$ gel set that has been used in a characterization of the N-terminal acetyl transferase NatB (Svensson and Blomberg, 2003) and (iii) a 1×12 gel set with technical replications of a single extract of yeast proteins. The protein labeling with ^{35}S-methionine, extraction and separation by 2-D gel electrophoresis were conducted as earlier described (Blomberg, 2002). The gel images were processed in the gel image software PDQuest (Bio-Rad Laboratories) with the following steps: image smoothing, background subtraction with a floating ball, spot detection, spot quantitation by 2-D Gaussian fitting, manual spot editing and manual spot matching. The sets contain

different strains, growth conditions, types of replication, gel running systems and image scanning devices. These differences are summarized in Table 1.

	2×5	2×2×3	1×12
Strain(s)	wild-type	wild-type *nat3*Δ	wild-type
Growth condition(s)	normal 1M NaCl	normal 1M NaCl	normal
Replications	5 biological	3 biological	12 technical
2nd dim. gel system	Investigator (Oxford Glycosystems)	Ettan DALT II (Amersham-Pharamacia)	Ettan DALT II (Amersham-Pharmacia)
Scanning device	Phosphoimager (Molecular Dynamics)	Phosphoimager (Bio-Rad Laboratories)	Phosphoimager (Bio-Rad Laboratories)
No. of matched spots	842	577	628

Table 1: The major differences between the three gel sets.

For each gel, the individual spot volumes are normalized by the total volume in all quantitated spots in the gel. This procedure accounts for pipetting errors when loading the samples to the first dimension strips, inconsistency in the transfer of proteins from the first dimension to the second dimension and different gel exposure times in the image acquisition. The normalized spot volume is given in parts per million, referred to as the original data throughout the paper and denoted with Z.

2.2 Variance versus mean stabilizing transformations

An initial exploratory analysis of the variation in normalized spot volumes using plots of the sample variance as a function of sample mean reveals a strong variance versus mean dependence among the different proteins. The dependence in the 2×5 gel set is shown in Fig. 1 (a) and (b) and indicates a multiplicative variation, where the dependence is quadratic and demonstrated as a line with slope two in plots with logarithmic scale on both axes.

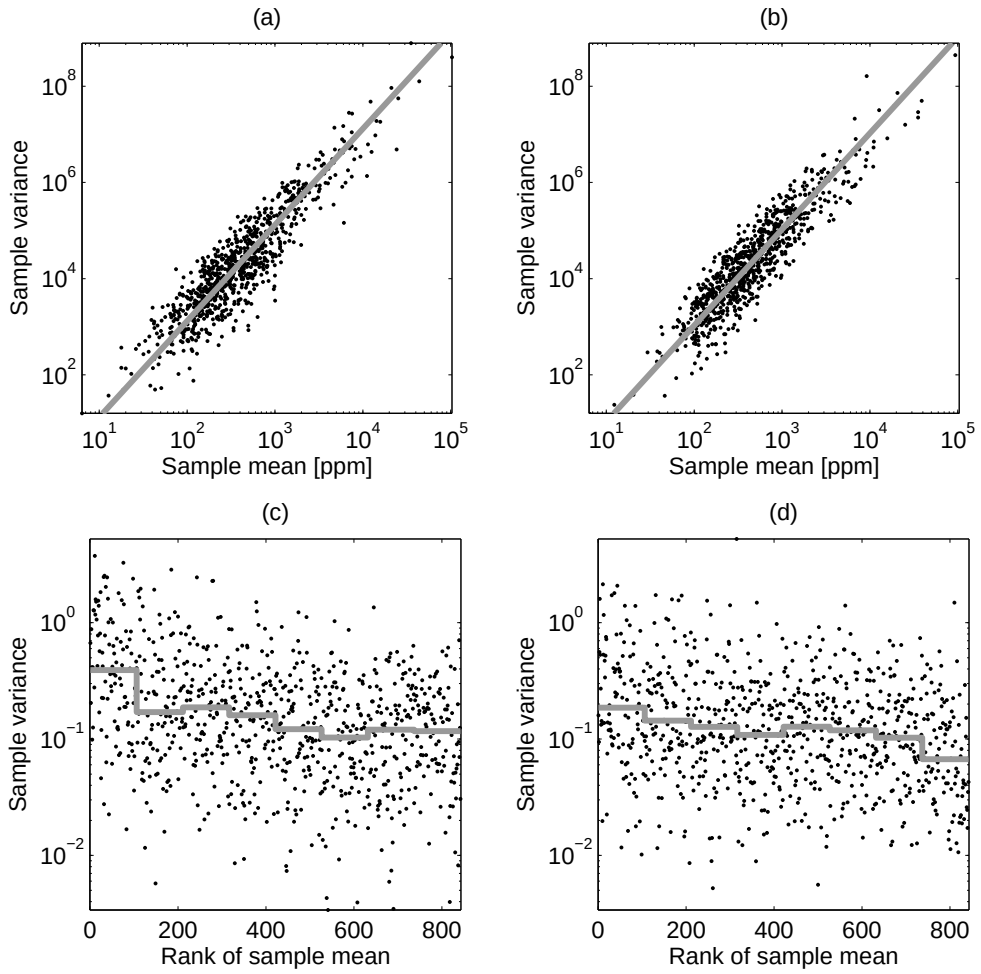

Figure 1: *First row.* Variance versus mean plots of original data in the 2×5 gel set: (a) normal and (b) 1M NaCl growth conditions. The gray lines in (a) and (b) have slope two; indicating a multiplicative error. *Second row.* Variance versus (rank of) mean plots of log transformed data: (c) normal and (d) 1M NaCl growth conditions. The gray lines in (c) and (d) show local medians. Logarithmic scales are used for both axes in (a) and (b) and for the vertical axis in (c) and (d). A dot in the figure corresponds to the measurements of one protein.

The standard technique to stabilize the variance of a stochastic variable with multiplicative variation is to apply a logarithmic transformation. However, we find that the logarithm is too strong for low abundance proteins. This can be seen in Fig. 1(c) and (d) where the variance of log transformed data (the natural logarithm is used throughout the paper) is plotted against the rank of the mean. Plotting against the rank distributes the data evenly along the x-axis and facilitates the visualization of the variance heterogeneity. The variance versus mean dependence is emphasized by plotting local medians: the x-axis is divided into eight intervals of equal length, the median of the sample variances in each interval is calculated and plotted as a horizontal line.

It may be noted that the standard deviation of log transformed data is approximately the same as the coefficient of variation (CV) of the original data, which in the literature is a common way to summarize variation in 2-D gel data. For example, by plotting and fitting a curve to the CV as a function of mean of spot volumes in 2-D gels with Coomassie Blue stained *Arabidopsis thaliana* proteins, Mahon and Dupree (2001) report a dependence similar to Fig. 1(c) and (d).

There are many sources of variation in 2-D gel data. The preceding exploratory analysis indicates that the variation mainly can be modeled with a multiplicative error but that this is not the whole story, especially in the case of low abundance proteins. A possible next step is to include an additive error component in the model.

2.2.1 The inverse hyperbolic sine transformation family

The inverse hyperbolic sine transformation family is motivated by a two-component model of measurement error (Rocke and Lorenzato, 1995) where the measured value is modeled as

$$Z = \zeta e^{\eta} + \varepsilon + \delta. \tag{1}$$

Here ζ is the true value, $\eta \sim N(0, \sigma_\eta^2)$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ are independent multiplicative and additive random error components, respectively, and δ is an additive bias.

The variance of data distributed according to model (1) is approximately stabilized by a transformation that is based on the inverse hyperbolic sine function (Huber et al., 2002; Durbin et al., 2002):

$$y_\lambda(z) = \operatorname{arsinh}\left(\frac{z + \lambda_2}{\lambda_1}\right). \tag{2}$$

This transformation holds for $\lambda_1 > 0$ and for all values of the numerator in the argument, even negative values. The relations between the parameters in the model and the approximate variance stabilizing transformation are $\lambda_1 \approx \sigma_\varepsilon/\sigma_\eta$ and $\lambda_2 \approx -\delta$. Hence, λ_1 reflects the additive (zero mean) random error and λ_2 reflects the additive bias.

Note that $\operatorname{arsinh}(x) = \log(x + \sqrt{x^2+1})$ for $x > 0$, so $\operatorname{arsinh}(x) \approx \log(2x)$ for large values of x. Consequently the shifted log transformation

$$y_\lambda(z) = \log(z + \lambda), \qquad z + \lambda > 0, \tag{3}$$

is, within an additive term, a limit case of the inverse hyperbolic sine transformations by letting $\lambda_1 \to 0$, *i.e.* when the additive (zero mean) random error is negligible.

2.2.2 Criterion for variance versus mean stabilization

A criterion for variance homogeneity and normality can be obtained by computing the log likelihood of original data under the assumption that for some unknown λ the transformed data are independently and normally distributed with equal variance (Box and Cox, 1964). With Z_{ijk} denoting the normalized spot volume of protein i, growth condition j and replication k, the assumption is that, independently for all i, j and k, $y_\lambda(Z_{ijk}) \sim N(\mu_{ij}, \sigma^2)$. For fixed λ, the maximized log likelihood with respect to the parameters μ_{ij} and σ^2 is

$$L_{max}(\lambda) = -\frac{1}{2}m(\log(2\pi)+1) - \frac{1}{2}m\log\hat{\sigma}^2(\lambda) + \sum_{ijk}\log\left|\frac{\partial y_\lambda}{\partial z}(z_{ijk})\right|, \tag{4}$$

where m is the total number of spots and $\hat{\sigma}^2(\lambda) = \sum_{ijk}(y_{ijk} - \bar{y}_{ij\cdot})^2/m$.

It is possible to factorize the log likelihood criterion into two parts (Box and Cox, 1964): (i) the contribution of the normality within each group of observations of a protein in a growth condition, and (ii) the contribution of variance homogeneity under assumption of normality. Usually, the information about λ coming from within group normality is very slight, whereas the requirement of constant variance has a major effect on the choice of λ. This is also the case for the examined 2-D gel data. Hence, the log likelihood criterion favors transformations that cause the group sample variances of transformed data to be homogeneous and consequently stabilize the variance versus mean dependence.

2.3 Exploration of transformed data

To further characterize the variation in the data, an analysis of variance of sample variances is performed after transformation and externally Studentized residuals are used to examine normality and to find spatial trends in the measurement errors.

2.3.1 Analysis of variance of sample variances

Analysis of variance models for the logarithm of the sample variances (Scheffé, 1959, pp 83-87) are used to analyze the dependence of sample variances on proteins, strains and growth conditions in the experimental design.

The 2×5 gel set is analyzed with the model

$$\log(s_{ij}^2) = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \tag{5}$$

where s_{ij}^2 is the sample variance of transformed data for protein i and condition j, α_i is the effect of protein i, β_j is the effect of growth condition j and ϵ_{ij} is the residual. While the growth condition effects should be regarded as fixed, the effects of proteins may be regarded either as fixed or random, depending on whether we are interested in estimating the individual protein effects or not.

The corresponding model for the 2×2×3 gel set is

$$\log(s^2_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijk}, \tag{6}$$

where γ_k denotes the effect of strain k, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ and $(\beta\gamma)_{jk}$ denote interactions, while the other components in the model have similar interpretations as in model (5).

2.3.2 Externally Studentized residuals

With $y_1, ..., y_n$ denoting the n transformed normalized spot volumes of a protein in a replicate group, the externally Studentized residuals are given by

$$t_i = \frac{y_i - \bar{y}}{s_{(i)}\sqrt{1 - 1/n}}, \qquad i = 1, ..., n, \tag{7}$$

where $\bar{y}$ is the sample mean and $s^2_{(i)}$ is the sample variance calculated by omitting the i:th observation. This cross-validation type of construction is standard in applied linear regression (see, for example, Draper and Smith (1998, p 208)) and ensures that under assumptions of normality and independence, the numerator and denominator are independent, and thus the externally Studentized residuals are t-distributed with $n - 2$ degrees of freedom.

Residuals are a useful tool to assess distributional assumptions and to diagnose systematic errors. This is because they are the best estimates of the random measurement errors. By dividing each residual by its estimated standard error, it is possible to compare residuals that correspond to different proteins, strains and growth conditions, and therefore might have different variances.

3 Results

3.1 Variance versus mean stabilization by transformation

To find the data transformation that stabilizes the variance versus mean dependence that can be seen in Fig. 1, we maximize the log likelihood criterion (4) with respect to the parameters in the inverse hyperbolic sine transformation (2). Contours of $L_{max}(\lambda)$, the log likelihood criterion as function of the two parameters in these transformations, are shown in Fig. 2. The estimates of the parameter λ_2 are $\hat{\lambda}_2 = 47$, 72 and 108 for the three data sets, and for all sets the maximum is obtained by letting $\lambda_1 \to 0$. Hence, the optimal transformation is the shifted logarithm (3). Figure 3 shows the result of applying the shifted log transformation to the data in the 2×5 gel set and is a variance versus mean stabilized counterpart of Fig. 1 (c) and (d).

The value of $L_{max}(\lambda)$ is not of interest in itself; it is the difference in log likelihood that provides information about the uncertainty in the parameter estimates and enables comparisons of different transformations. For example, the increment in log likelihood from the ordinary logarithm, that is given by letting $\lambda_1 \to 0$ and $\lambda_2 = 0$, to the optimal shifted

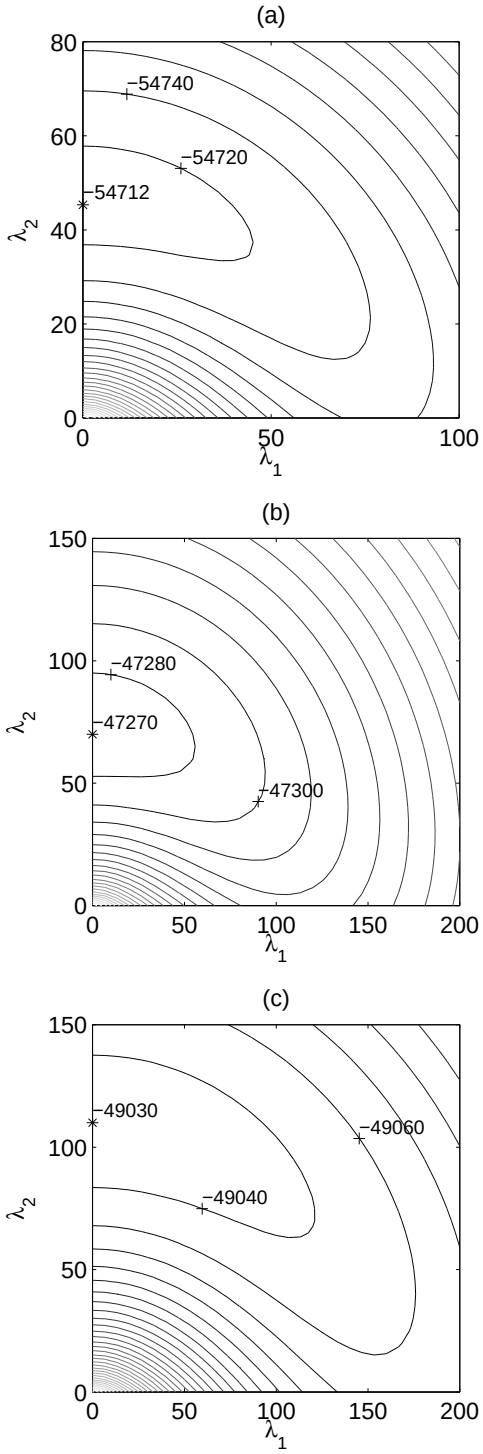

Figure 2: Contours of $L_{max}(\lambda)$: (a) the 2×5 gel set, (b) the 2×2×3 gel set and (c) the 1×12 gel set. The step between consecutive shown contours is 20 units in the log likelihood. The location of maxima are marked with a star and the corresponding maximal values are shown.

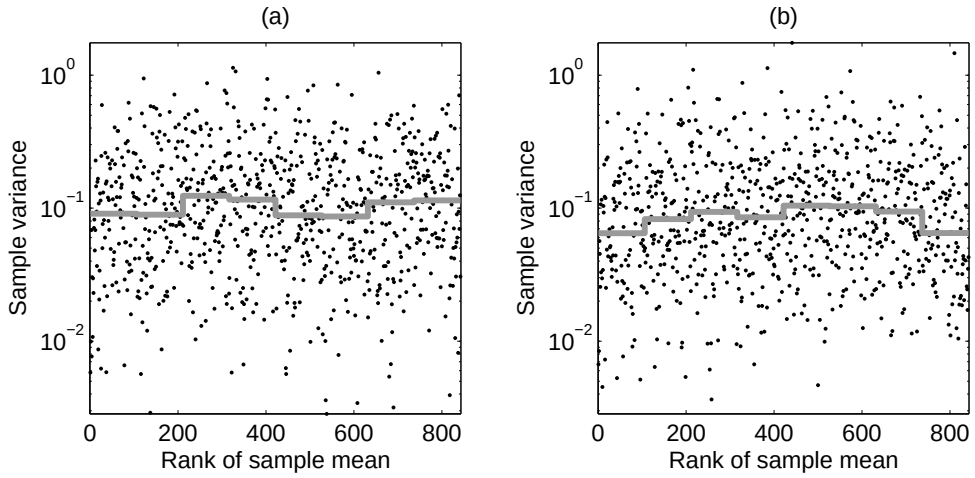

Figure 3: Variance versus (rank of) mean plots of $\log(z+47)$ transformed data in the 2×5 gel set: (a) normal and (b) 1M NaCl growth conditions. Gray lines show local medians.

log transformation is highly significant in all three data sets. The ridges of the surfaces in Fig. 2 can be understood with the concept of power strength of a transformation (Tukey, 1957). The transformations along the ridge have similar power strength, *i.e.* they are comparable, for values in the bulk of the distribution of data.

It may be noted that we also tried the power transformation family formulated by Box and Cox (1964): $y_\lambda(z) = ((z+\lambda_2)^{\lambda_1}-1)/\lambda_1$. Compared to the shifted log transformation it gives a marginal but statistically significant improvement. The estimates of the exponent λ_1 are $\hat{\lambda}_1 = 0.04$, 0.11 and 0.06 for the three data sets. However, as the improvement is small, our conclusion is that the conceptually simpler shifted log transformation seems adequate for further analysis.

3.2 Exploration of transformed data

3.2.1 Variance heterogeneity among proteins

Having stabilized the variance versus mean dependence in data, we proceed by examining if, although not depending on the mean, there is a variance heterogeneity in transformed data and try to determine the major source of such a heterogeneity. This is done with analysis of variance models of the logarithm of the sample variances. Tables 2 and 3 are the ANOVA tables for the models (5) and (6) for the 2×5 and the $2\times2\times3$ gel sets, respectively.

Booth ANOVA tables show a clear variance heterogeneity for different proteins and the empirical distributions of the estimated protein effects are shown in Fig. 4. No interactions are significant in Table 3, which simplifies the interpretation. There are minor significant effects of the growth conditions in both tables, but these effects are small compared to

Analysis of variance

Source	Sum Sq.	d.f.	Mean Sq.	F
Protein	1241.3	841	1.48	2.07^{***}
Condition	4.5	1	4.46	6.28^{*}
Error	598.3	841	0.71	
Total	1844.1	1683		

Table 2: ANOVA table of model (5) for the 2×5 gel set. In the column with F-statistics (***) denotes $p < 0.001$ and (*) denotes $p < 0.05$.

Analysis of variance

Source	Sum Sq.	d.f.	Mean Sq.	F
Protein	1393.3	576	2.42	1.42^{***}
Condition	9.2	1	9.23	5.43^{*}
Strain	1.8	1	1.83	1.08
Protein×Condition	1074.0	576	1.86	1.10
Protein×Strain	1012.3	576	1.76	1.03
Condition×Strain	0.1	1	0.08	0.05
Error	978.9	576	1.70	
Total	4469.7	2307		

Table 3: ANOVA table of model (6) for the 2×2×3 gel set.

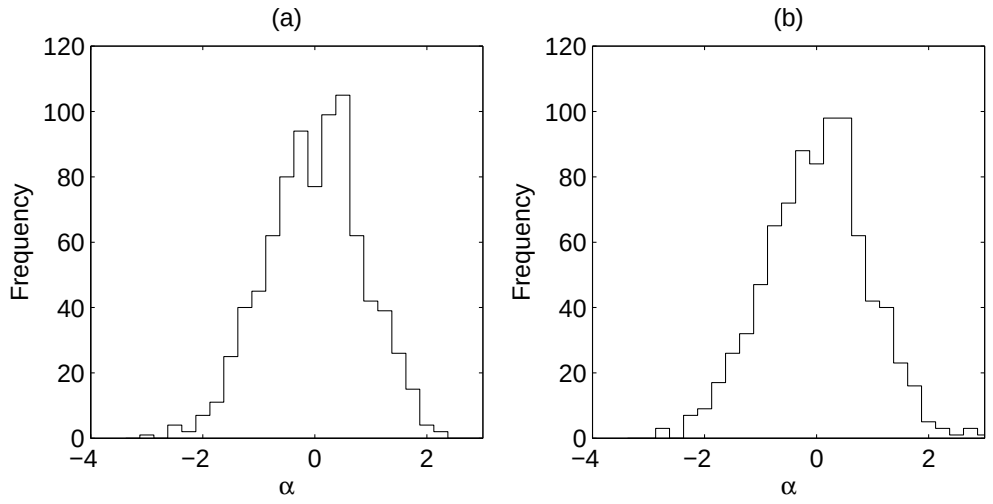

Figure 4: Empirical distribution of estimates of the protein effects on the logarithm of sample variances: (a) in model (5) for the 2×5 gel set, and (b) in model (6) for 2×2×3 gel set.

the protein effects. Hence, the difference in precision in the measurement of two randomly chosen proteins in a gel, is on average larger than the difference in precision in the measurement of the same protein in two gels from different strains and/or growth conditions.

In the analysis of variance, residuals are assumed independent, which seems quite reasonable for different strains and growth conditions, but more arguable for different proteins. The result of the analyses of variance is therefore regarded as exploratory rather than conclusive.

3.2.2 Residual plots reveal approximate normality and spatial trends

Finally, we explore the residuals – the estimates of the random measurement errors. Since we have found a remaining variance heterogeneity for different proteins, we work with the externally Studentized residuals (7). Here we divide each residual by its estimated standard error which enables a comparison of the residuals that correspond to different proteins.

To examine the issue of normality, we regard the marginal distribution of all externally Studentized residuals in a replicate group. Histograms of these residuals in the 2×5 gel set are shown in Fig. 5 together with the probability density of a t-distributions with 3 degrees of freedom corresponding to the assumption of independent and normally distributed errors in transformed data. The empirical distributions are somewhat flattened in the center and have slightly heavier tails, but the deviations do not seem alarming. Similar observations are made for the other two data sets (data not shown).

The residuals are further used to demonstrate that the measurements errors are spatially correlated. Figure 6 shows the externally Studentized residuals in the 2×5 gel set as function of spot positions in the gel. There are areas in the gels where the spots volumes

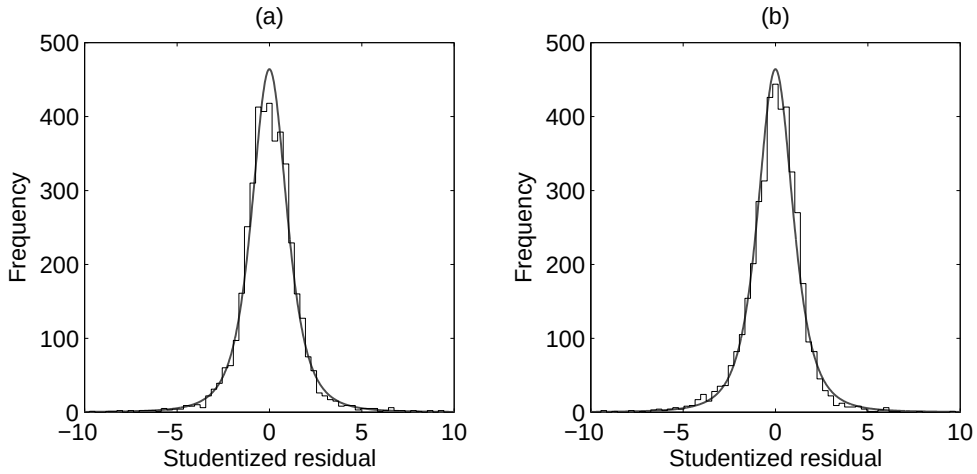

Figure 5: Histograms of externally Studentized residuals in the 2×5 gel set: (a) normal and (b) 1M Na Cl growth conditions. The smooth curves show the t-distribution with 3 degrees of freedom.

generally are higher or lower than the average of the values in the replicate group. Similar spatial trends can be observed in the other two gel sets (data not shown). To understand the size of the spatial trends, we perform the following quick calculation. If we consider a typical protein with a CV of 35% and disregard the additive bias, then an externally Studentized residual of 1.5 for a protein spot in one gel roughly corresponds to a 1.5-fold spot volume change relative to the (geometric) mean of the spot volumes in the replicate group.

4 Discussion

4.1 Variance versus mean stabilizing data transformation

The estimated transformation parameters can be viewed in two ways: (i) the transformations are merely seen as tools to obtain variance versus mean stabilization for a subsequent statistical analysis and the estimated parameters are not assigned further interpretation; or (ii) confidence is placed in the two-component model (1), and with the assumption that the variance should not be dependent on the mean, the estimated transformation parameters can be seen as estimates of parameters in the model. Hence, $\hat{\lambda}_1 \approx 0$ indicates that the additive (zero-mean) random error is negligible and $\hat{\lambda}_2 > 0$ suggests that there is a negative additive bias. If the estimated $\hat{\lambda}_2$ is considered a small value that can be attributed to uncertainty in the measurement procedure, then $z + \hat{\lambda}_2$ might, in the latter point of view, be seen as a (statistically) calibrated value of the observed normalized spot volume z similar to the simultaneous calibration and variance stabilization of microarray data proposed by Huber et al. (2002).

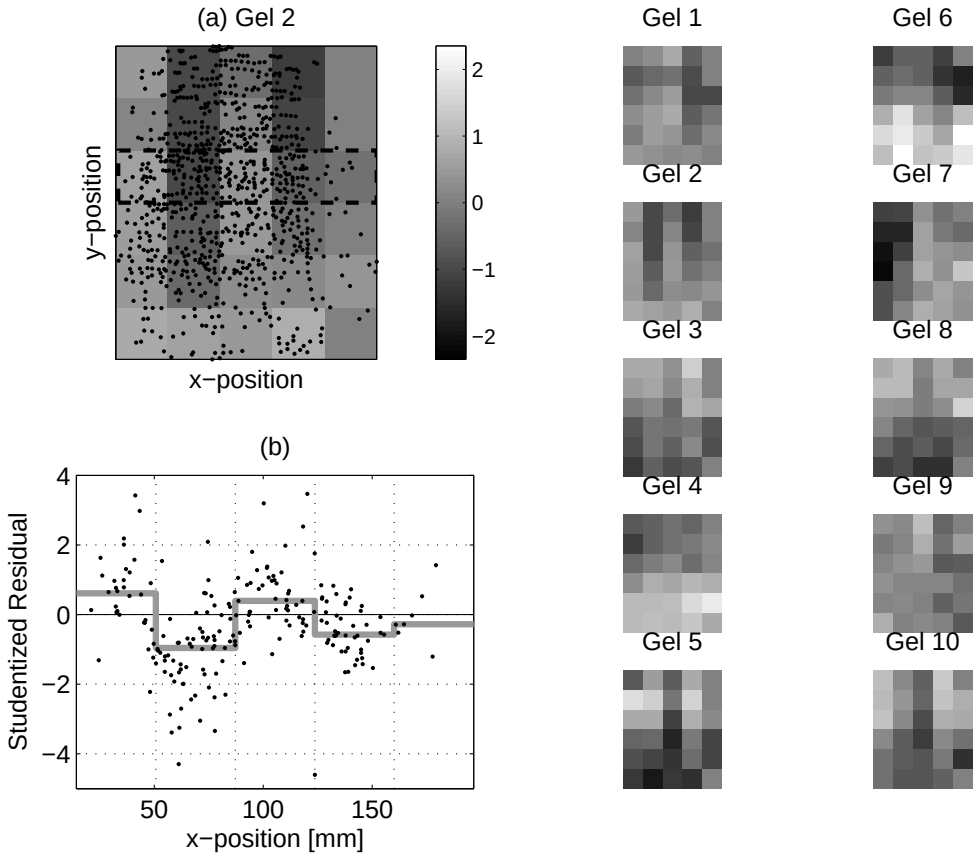

Figure 6: (a) The spatial correlations among the externally Studentized residuals in the second gel in the 2×5 gel set are illustrated with local medians according to the gray scale in the bar to the right. The dots mark spot positions. The residuals in the third row of local medians that is marked in (a) are plotted in (b) as a function of the x-position and the local medians are shown with gray lines. The spatial trends in all ten gels are shown to the right in the figure: gels 1-5 are from normal growth conditions and gels 6-10 are from 1M NaCl growth conditions.

4.2 Variance heterogeneity among proteins

Even after use of a shifted logarithmic transformation to compensate for the variance versus mean dependence, there remains a clear variance heterogeneity for different proteins as shown by the analysis of variance in Sec. 3.2.1. We have investigated if this variability is related to the isoelectric point (x position) or size (y position) of the proteins, but we do not found any clear such connections (data not shown). It remains to further analyze this heterogeneity and, for example, to see how stable it is when we compare different experiments, and if there are some specific properties of the proteins with high or low variability.

4.3 Spatial trends

The image background can vary in different gel images, either globally over the whole or large areas of the gel or locally around specific spots. The customary approach in gel image processing is to rely on extrapolation of image background variability to the spot areas to estimate the background level in the spot. Hence, a poor background estimation in the image processing might contribute to spatial correlations among quantitated spot volumes. It may be noted that the background subtraction is still an issue of discussion in the 2-D gel community, for example Mahon and Dupree (2001) have reported reduced reproducibility in Coomassie Blue stained protein gels when the image background is estimated and subtracted as compared to no subtraction.

There might also be biological phenomena that can manifest as spatial trends. An example is proteolysis – the enzymatic fragmentation and degradation of proteins. A higher degree of proteolysis will decrease the volumes of spots corresponding to full-size proteins, and increase the quantities of the spots corresponding to their degradation fragments that are positioned further down the second dimension in the gel. This is hypothesized as being the situation in the $2\times2\times3$ gel set, where the 2-D gel data indicate that the *nat3*Δ mutant has an increased general protease activity and that this may be due to decreased activity of a protease inhibitor that is dependent on N-terminal acetylation (Svensson and Blomberg, 2003).

It is desirable to correct for systematic errors in terms of spatial trends and thus improve the measurement precision. Dowsey et al. (2003) have applied a method from the magnetic resonance imaging field (Lai and Fang, 1999) to correct for regional relative bias in pairs of gel images. The method corrects for very slowly varying trends in ratios of quantitated expression between the two spots in each pair-wise match. We think one should be aware of the risk that systematic spatial trends is confounded with the effect of a biological phenomenon such as proteolysis, and by normalizing spatial trends in data using a between treatment group comparison, one might conceal biologically relevant information.

In the design of experiments, the proper procedure to avoid the confusion of relevant information and systematic errors is to randomize. In the case of spatial trends this would correspond to a randomization of the positions of the protein spots in the gel, similar to the randomization of gene probe positions in the design of microarrays. However, this is

simply not possible in the case of 2-D gels; a protein spot position is determined by the physiochemical properties of the protein and the position is roughly the same in all gels, which is the very foundation of the 2-D gel technology.

We see the spatial correlations together with the small additive bias as a motivation to investigate the image analysis step – especially the image background estimation and subtraction. The statistical techniques described in this paper may serve as quality measures in such an investigation.

4.4 On the statistical analysis of 2-D gel data

In the exploratory analyses we find that the transformed data are approximately normally distributed with a clear variance heterogeneity among different proteins. Hence, for the variance versus mean stabilized data and with awareness of the spatial correlations, we can apply the machinery of statistical methods that are based on normality. For example, these observations justify the use of individual two-sample t-tests for differential expression. Instead of comparing the t-statistics to quantiles of the corresponding t-distribution, it is possible to use resampling methods, *e.g.* bootstrap or permutation techniques, to estimate the unknown null distribution of the test statistics and thus avoid parametric assumptions. For example, Dudoit et al. (2003) have used the permutation technique in multiple hypothesis testing for differential expression in microarray data. A permutation of the gels in the gel set would create a situation where the correlation structure, for example due to the spatial correlations, among measured protein levels are preserved. However, the number of gels in the gel sets that are considered in this paper is on the border of being too small for resampling methods. In computer simulations using normally distributed random variables, Dudoit et al. (2003) found that for 500 genes and the small sample size of five (as in the 2×5 gel set), the result of a permutation procedure is more conservative than using quantiles of the corresponding t-distribution.

Individual t-tests are probably too precautionary – they guard against any distribution of variance among proteins, and as indicated in Fig. 4, the underlying distribution of variances does not seem to be too irregular. A possible approach is to use a hierarchical model with an inverse gamma distribution for the variances and normal distributions for the observations given the variances. This is the hierarchical model that has been used by Baldi and Long (2001) and Lönnstedt and Speed (2002) for Bayesian inference from microarray data. We have applied this model to our 2-D gel data in the following way: (i) similarly to Lönnstedt and Speed (2002), we use the empirical Bayes technique of estimating the hyper-parameters in the inverse gamma distribution from data; and (ii) in the spirit of Baldi and Long (2001), we take a short cut from a full Bayesian analysis and regularize the ordinary two-sample t-statistics by replacing the sample variances with empirical Bayes variance estimates. It may be noted that the result of using these regularized t-statistics for the transformed data in our 2-D gel sets is actually quite similar to the result in the combination of ordinary two-sample t-tests of log transformed data and a fold change criterion. This method is commonly used in the analysis of 2-D gel data (see, for example, Blomberg (2002)), and requires the specification of the significance level in the individual

t-tests as well as a specification of the *ad hoc* fold change threshold. The latter specification can be avoided using regularized t-tests. Note that we have here disregarded the multiple comparison issue in the hypothesis testing.

Acknowledgments

This work has been financially supported by the Bioscience Program at Chalmers University of Technology.

References

Alipour, H., Farea, T., Ericson, E., Mostad, P., Norbeck, J., Nerman, O. and Blomberg, A. (2003). Steady-state salt stress response in *Saccharomyces cerevisiae*: relation to protein expression and functional importance. Submitted.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes, *Bioinformatics* **17**: 509–519.

Blomberg, A. (2002). Use of two-dimensional gels in yeast proteomics, *Methods in Enzymology* **350**: 559–584.

Blomberg, A., Blomberg, L., Norbeck, J., Fey, S., Larsen, P., Larsen, M., Roepstorff, P., Degand, H., Boutry, M., Posch, A. and Görg, A. (1995). Interlaboratory reproducibility of yeast protein patterns analyzed by immobilized pH gradient two-dimensional gel electrophoresis, *Electrophoresis* **16**: 1935–1945.

Box, G. and Cox, D. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**: 211–246.

Burstin, J., Zivy, M., de Vienne, D. and Damerval, C. (1993). Analysis of scaling methods to minimize experimental variations in two-dimensional electrophoresis quantitative data: Application to the comparison of maize inbred lines, *Electrophoresis* **14**: 1067–1073.

Dowsey, A., Dunn, M. J. and Yang, G.-Z. (2003). The role of bioinformatics in two-dimensional gel electrophoresis, *Proteomics* **3**: 1567–1596.

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*, John Wiley & Sons.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science* **18**: 71–103.

Durbin, B., Hardin, J. S., Hawkins, D. and Rocke, D. (2002). A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics* **18**: S105–S110.

Garrels, J. I. (1989). The QUEST system for quantitative analysis of two-dimensional gels, *Journal of Biological Chemistry* **264**: 5269–5289.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics* **18**: S96–S104.

Lai, S. H. and Fang, M. (1999). A new variational shape-from-orientation to correct intensity inhomogeneities in magenetic resonance images, *Medical Image Analysis* **3**: 409–424.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data, *Statistica Sinica* **12**: 31–46.

Mahon, P. and Dupree, P. (2001). Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full, *Electrophoresis* **22**: 2075–2085.

Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T. and VanBogelen, R. A. (2003). Overcoming technical variation and biological variation in quantitative proteomics, *Proteomics* **3**: 1912–1919.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins, *Journal of Biological Chemistry* **250**: 4007–4021.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (eds) (2003). *The Analysis of Gene Expression Data*, Springer.

Rocke, D. and Lorenzato, S. (1995). A two-component model for measurement error in analytical chemistry, *Technometrics* **37**(2): 176–184.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley.

Speed, T. P. (ed.) (2003). *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall.

Svensson, R. and Blomberg, A. (2003). The study of protein modifications in *Saccharomyces cerevisiae*: Characterization of the N-terminal acetyl transferase NatB, Licentiate thesis, Göteborg University.

Tukey, J. W. (1957). On the comparative anatomy of transformations, *Annals of Mathematical Statistics* **28**: 602–632.

Voss, T. and Haberl, P. (2000). Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis, *Electrophoresis* **21**: 3345–3350.