*Submitted to the Annals of Statistics*

# DISCUSSION OF "A SIGNIFICANCE TEST FOR THE LASSO"

By Peter Bühlmann, Lukas Meier and Sara van de Geer

*ETH Zürich*

We congratulate Richard Lockhart, Jonathan Taylor, Ryan Tibshirani and Robert Tibshirani for a thought provoking and interesting paper on the important topic of hypothesis testing in potentially high-dimensional settings.

**1. A short description of the test procedure.** We start by presenting the proposed test procedure in a slightly different form than in the paper. Let

$$\hat{\beta}(\lambda) := \arg\min \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

be the Lasso estimator with tuning parameter equal to $\lambda$. The paper uses the Lasso path $\{\hat{\beta}(\lambda) : \lambda > 0\}$ to construct a test statistic for the significance of certain predictor variables.

For a subset $S \subseteq \{1, \ldots, p\}$ let $\hat{\beta}_S(\lambda)$ be the Lasso solution using only the variables in $S$:

$$\hat{\beta}_S(\lambda) := \arg\min_{\beta_S \in \mathbb{R}^{|S|}} \frac{1}{2}\|y - X_S\beta_S\|_2^2 + \lambda\|\beta_S\|_1.$$

The covariance test is based on the difference

$$T(S,\lambda) := \Big[\|y - X_S\hat{\beta}_S(\lambda)\|_2^2 + \lambda\|\hat{\beta}_S(\lambda)\|_1\Big]/\sigma^2 - \Big[\|y - X\hat{\beta}(\lambda)\|_2^2 + \lambda\|\hat{\beta}(\lambda)\|_1\Big]/\sigma^2.$$

If $T(S,\lambda)$ is large then the solution using only the values in $S$ does not have a very good fit and this may support evidence against the hypothesis $H_S: \ A^* \subseteq S$, where $A^* = \text{support}(\beta^*)$ is the true active set.

Let $\infty =: \hat{\lambda}_0 > \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$ be the knots of $\hat{\beta}(\lambda)$. For $k \geq 1$, let $\hat{A}_k := \text{support}(\hat{\beta}(\hat{\lambda}_k))$. We put "hats" on these quantities to stress that they are random variables depending (only) on the data. Thus $T(\hat{A}_k, \hat{\lambda}_k) = 0$ and by continuity arguments also $T(\hat{A}_{k-1}, \hat{\lambda}_k) = 0$. The authors suggest to use the test statistic

$$T_k := T(\hat{A}_{k-1}, \hat{\lambda}_{k+1})$$

for the hypotheses $H_{\hat{A}_{k-1}}$. They derive the interesting result that under certain conditions, the test statistic has an asymptotic exponential distribution.

**2. A "conditional" test.** Fixing the value of $k$, the test is a conditional test for $H_S$ given that $\hat{A}_{k-1} = S$ (the event one conditions on is denoted below and in the paper by $B$; the paper presents two versions, in Section 3.2 and 4.2, respectively). Such kind of a test is uncommon: the usual form of a conditional test is to condition on an observable event, for example when conditioning on an ancillary statistics (Ghosh et al., 2010, cf.). Here, however, the conditioning event $B$, that all active variables enter the Lasso path first, is *unobserved*.

The difficulty with such an unobserved event is treated in the paper by imposing sufficient conditions such that $\mathbb{P}[B] \to 1$ asymptotically, and therefore, one can simply ignore the effect of conditioning. The imposed conditions are rather restrictive: in particular, they include a "beta-min" assumption requiring that the nonzero regression coefficients are sufficiently large in absolute value. We illustrate in Figure 1 that the lower bound for the non-zero coefficients (beta-min) has to be large or very large in order that the active set is correctly identified right at the first steps of the Lasso path (the latter is the conditioning event $B$ as in Section 3.2 of the paper while in Section 4.2 of the paper, a slightly different version of $B$ is presented; we believe that the quantitative differences in terms of $P(B)$ are small). Based on this observation, we imagine that the obtained limiting distribution in Theorems 1 and 3 often does not approximately capture the conditional distribution of the test statistics (when conditioning on the event $B$), and there is no strong guarantee that the obtained p-values would be approximately correct in practical settings. It would be interesting to work out a correction factor which would take into account that $\mathbb{P}[B]$ is not close to one: we do not know how this could be achieved.

2.1. *Interpretation of the p-values.* A correct interpretation of the proposed p-values from the covariance test seems not straightforward. First, these p-values are *not* justified to lead to significance statements for fixed variables (or hypotheses) since the test is a conditional test. For example, for the wine data in the right column of Table 5 in the paper, the p-value for the variable "`pH`" should not be interpreted in the classical sense based on a fixed null-hypothesis $\beta^*_{\mathrm{pH}} = 0$. In many scientific applications and contexts, such classical p-values are desired (maybe after adjustment for multiple testing), and we think that the covariance test does not really provide such p-values; in fact, the authors never make such a claim. Reasons for the statement above include: (i) the covariance test only assigns significance of "the $k$th variable entering the Lasso path", but since the $k$th variable is random (possibly even when $\mathbb{P}[B] \to 1$), there seems to be an issue to

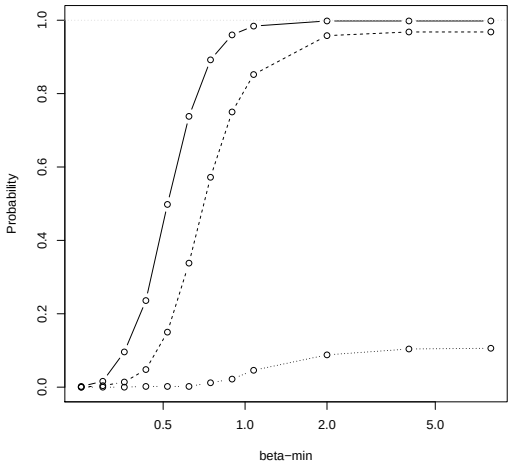


FIG 1. *Empirical probabilities (500 simulation runs) for the event that all truly active coefficients are identified in the first $k_0$ steps of the Lasso path. A Gaussian AR(1) design matrix with $n = 100, p = 1000$ and $\rho = 0.5$ is used (i.e., the population covariance matrix $\Sigma$ is Toeplitz with $\Sigma_{ij} = \rho^{|i-j|}$). The active set has size $k_0 = 3$ (solid), $k_0 = 5$ (dashed) and $k_0 = 10$ (dotted). The active coefficients are placed at random positions and all coefficients are of the same size (beta-min). The error variance $\sigma^2$ is set to 1.*

map the $k$th variable to a fixed variable, such as "`pH`" or "`alcohol`"; (ii) in view that $\mathbb{P}[B]$ might be far away from one as illustrated in Figure 1, the interpretation should be conditional, i.e., "given that all active variables enter the Lasso path first"; and such a conditional interpretation of a p-value seems somewhat awkward. We briefly outline in Section 5 some alternative methods which are mathematically justified for classical (fixed hypotheses) p-values in a high-dimensional context.

Our question to the authors is how to interpret the p-values in practice. In view of available software, there is a substantial risk that practitioners blindly use and interpret the obtained p-values as usual (for fixed hypotheses), and hence, some guidance for proper use and interpretation would be very useful.

**3. The assumptions.** The authors require a condition on the design matrix and a beta-min assumption. These assumptions are used to guarantee that the conditioning event $B$, namely that the first $k_0$ variables entering the Lasso path contain the active set, has large probability.

In Theorem 3 of the paper, an irrepresentable condition (Zhao and Yu,

2006) is assumed. Let $A_0 \supseteq A^*$ and let

$$\eta > \max_{j \notin A_0} \sup_{\|\tau_{A_0}\|_\infty \leq 1} |X_j^T X_{A_0} (X_{A_0}^T X_{A_0})^{-1} \tau_{A_0}|.$$

We assume the irrepresentable condition $\eta \leq 1$. From Exercise 7.5 in Bühlmann and van de Geer (2011) we know that for $\lambda_\eta := \lambda_\epsilon (1+\eta)/(1-\eta)$ we have $\hat{A}(\lambda_\eta) \subseteq A_0$. Here

$$\lambda_\epsilon = \max_{1 \leq j \leq p} | < \epsilon, X_j > |.$$

Define now

$$\hat{k}_\eta := \max\{k : \lambda_k \geq \lambda_\eta\}.$$

Thus with large probability

$$A_* \subseteq \hat{A}_{\hat{k}_\eta} \subseteq A_0.$$

We imagine moreover that in practice one would follow the Lasso path and steps as soon as the test accepts $\hat{A}_{k-1}$. Define therefore $\hat{k}$ as being the first $k$ for which the hypothesis $H_{\hat{A}_{k-1}}$ is accepted. In line with the paper, one then assumes $\hat{A}_{\hat{k}-1} \supseteq A_0$, and then with probability approximately $1-\alpha$, $\hat{A}_{\hat{k}-1} = A_0$. Alternatively, applying this argument to $\hat{A}_{\hat{k}_\eta}$ (which is allowed since $A_* \subseteq \hat{A}_{\hat{k}_\eta}$) we get $\hat{k}_\eta \geq \hat{k} - 1$ with probability approximately $1-\alpha$ and then we end up with $A_* \subseteq \hat{A}_{\hat{k}-1} = \hat{A}_{\hat{k}_\eta} = A_0$.

A related screening property of the Lasso is known (Bühlmann and van de Geer, 2011, cf. Ch.2.5): for $\lambda \asymp \sqrt{\log(p)/n}$,

$$\mathbb{P}[\hat{A}(\lambda) \supseteq A^*] \to 1, \tag{1}$$

assuming the compatibility condition on the design and a beta-min assumption. We note that the compatibility condition is weaker than the irrepresentable condition mentioned above (van de Geer and Bühlmann, 2009).

The authors argue in their Remark 4 that the beta-min assumption can be relaxed. Such kind of a relaxation is given in Bühlmann and Mandozzi (2013), assuming a zonal assumption allowing that some but not too many non-zero regression coefficients are small. It is also shown that zonal assumptions are necessary for validity of a sampling splitting procedure (Wasserman and Roeder, 2009), and we believe that a justification of the covariance test also necessarily needs some version of zonal assumptions. We remark that "in practice", achieving a statement as in (1) or saying that $\mathbb{P}[B] \approx 1$ (as in the paper) seems often unrealistic, as illustrated in Figure 1 and in Bühlmann and Mandozzi (2013).

3.1. *Hypothesis testing and assumptions on $\beta^*$.* In view of the fact that assumptions about $\beta^*$ are (have to be) made, the covariance test is exposed to the following somewhat undesirable issue. A significance test should *find out* whether a regression coefficient is sufficiently large. Thus, a zonal or beta-min assumption rules out the *essence of the question* by assuming that most or all non-zero coefficients are large. We note that (multi) sample splitting techniques (Wasserman and Roeder, 2009; Meinshausen et al., 2009) for hypothesis testing in high-dimensional scenarios suffer from the same problem. The procedure outlined in Section 5 does not make such zonal or beta-min assumptions.

**4. The power of the covariance test.** The paper does not make any claim about the power of the test nor does it include a comparison with other methods; regarding the latter, see Section 5.1.

Under the beta-min assumption, a theoretical study of the test's power is uninteresting: asymptotically, the power of the test is approaching one. Non-trivial power statement require that the non-zero regression coefficients are in the $1/\sqrt{n}$ range but this is excluded by the imposed beta-min assumption.

The following thoughts might lead to some insights for which scenarios the covariance test is expected to perform (reasonably) well. In an alternative and simplified setup, one could think of using a refitting procedure to test significance. Let

$$\hat{\beta}_S := \hat{\beta}_S(0) = \arg\min_{\beta_S \in \mathbb{R}^{|S|}} \|y - X_S \beta_S\|_2^2,$$

and for $\tilde{S} \supseteq S$

$$\begin{aligned} T(S, \tilde{S}) &:= \|y - X_S \hat{\beta}_S\|_2^2/\sigma^2 - \|y - X_{\tilde{S}} \hat{\beta}_{\tilde{S}}\|_2^2/\sigma^2 \\ &= \Big( < y, X_{\tilde{S}} \hat{\beta}_{\tilde{S}} > - < y, X_S \hat{\beta}_S > \Big)/\sigma^2. \end{aligned}$$

An alternative test statistic would then be $T(\hat{A}_{k-1}, \hat{A}_k)$. In the case of orthogonal design, we get

$$T_k = (\hat{\lambda}_k^2 - \hat{\lambda}_k \hat{\lambda}_{k+1})/\sigma^2, \quad T(\hat{A}_{k-1}, \hat{A}_k) = \hat{\lambda}_k^2/\sigma^2. \tag{2}$$

Obviously, if we fix $S$ and $j \notin S$, we get $T(S, S \cup \{j\}) = (< y, X_j >)^2/\sigma^2$ which has under $H_S$ a $\chi^2(1)$ distribution. If $\mathbb{P}(\hat{A}_{k-1} \supseteq A^*) \to 1$, then for each $j \notin \hat{A}_{k-1}$, $T(\hat{A}_{k-1}, \hat{A}_{k-1} \cup \{j\})$ is asymptotically $\chi^2(1)$. However $T_k$ and $T(\hat{A}_{k-1}, \hat{A}_k)$ are tests where the the decision which variable is to be tested for significance depends on the data. For the case of orthogonal design and

$A^* = \emptyset$, we have $\mathbb{P}_{H_\emptyset}(\hat{A}_{k-1} \supseteq A^*) = 1$, and $T(\hat{A}_{k-1}, \hat{A}_k)$ is approximately distributed as the $k$-th order statistic of a sample from a $\chi^2(1)$-distribution in decreasing order. For $k = 1$ (say) the statistic $T_1$ has a different scaling under the hypothesis $H_\emptyset : \ A^* = \emptyset$ because the order statistics behave like

$$T(\emptyset, \hat{A}_1) = \mathcal{O}_{\mathbb{P}_{H_\emptyset}}(\log n)$$

($p = n$ in the orthonormal case) whereas $T_1$ has asymptotically an exponential distribution, a nice fact proved in the paper, so that $T_1 = \mathcal{O}_{\mathbb{P}_{H_\emptyset}}(1)$. This means that $T_1$ has more power to detect alternatives of the form $H_{\{j\}} : \ A^* = \{j\}$. But it may have less power for alternatives of the form $H_{\{j_1,j_2\}} : \ A^* = \{j_1, j_2\}$. Under this alternative, $A^* \neq \hat{A}_0$ and if the two non-zero coefficients are very close together it will downscale the statistic $T_1$. This can also be seen from the expression (2): if the non-zero coefficients are similar, then

$$\hat{\lambda}_{k-1} \approx \hat{\lambda}_k$$

which leads to small values for $T_k$ while this has no (substantial) effect on $T(\hat{A}_{k-1}, \hat{A}_k)$: thus, the covariance test might be sub-ideal for detection of coefficient vectors whose individual non-zero coefficients are similar (as in the simulated examples in the paper and in Section 5.1). It would be interesting to better understand the regimes where the covariance test has strong and weak power.

**5. Alternative methods.** Other methods leading to p-values for fixed hypotheses $H_{0,j} : \ \beta_j^* = 0$ have been proposed in earlier work (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Minnier et al., 2011; Bühlmann, 2013; Chatterjee and Lahiri, 2013; Zhang and Zhang, 2011). We outline here the method from Zhang and Zhang (2011) which has been further analyzed in van de Geer et al. (2013) and Javanmard and Montanari (2013). The idea is to de-sparsify the Lasso, resulting in a new estimator $\hat{b}$ which is not sparse. Due to non-sparsity, this new $\hat{b}$ will not be suitable for prediction in high-dimensional settings, but its $j$th component $\hat{b}_j$ is asymptotically optimal for the low-dimensional target $\beta_j^*$ of interest:

$$\sqrt{n}(\hat{b}_j - \beta_j^*) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 v_j), \tag{3}$$

where $\sigma_\varepsilon^2 v_j$ is the Cramer-Rao lower bound. Such a result needs some assumptions on the design and sparsity of $\beta^*$ but no further restrictions on $\beta^*$ in terms of zonal or beta-min assumptions (van de Geer et al., 2013; Javanmard and Montanari, 2013). Thus, we are in the semiparametric framework,

where we can optimally estimate a low-dimensional parameter of interest in presence of a very high-dimensional nuisance parameter $\eta = \{\beta_k^*;\ k \neq j\}$: notably, we have the $1/\sqrt{n}$ convergence rate, even when $p \gg n$, and the best possible constant in the asymptotic variance.

The analysis in van de Geer et al. (2013) also shows that (3) holds *uniformly* over all sparse parameter vectors $\beta^*$ and therefore, the obtained confidence intervals and tests are honest. This is not the case when using a residual-based bootstrap in Chatterjee and Lahiri (2013) which exhibits the unpleasant super-efficiency phenomenon. As a consequence, post-model selection techniques (Leeb and Pötscher, 2003; Berk et al., 2013) are not necessary to construct valid, and in fact most powerful, hypothesis testing.

5.1. *A small empirical comparison.* We present here some result from a small simulation study based on a similar model as the Gaussian $AR(1)$ model in the paper with $\rho = 0.5$. We use an active set $A^*$ of size 10, where the active coefficients are placed at random positions and all have the same size. A total of 500 simulation runs are performed for each scenario.

We consider two-sided testing of individual hypotheses $H_{0,j} : \beta_j^* = 0$, possibly with adjustment for multiple testing using the Bonferroni-Holm procedure to control the familywise error rate.

The covariance test is used in the following two ways. A first approach (denoted by cov) is to follow the Lasso path until the first time the (unadjusted) p-value of the covariance test is non-significant and declare all corresponding predictor variables as significant which entered before such a non-significance flag of the covariance test. A second approach (denoted by cov.pval) is to assign those predictors, that remain in the Lasso path until the end, the p-value of the covariance test when they last entered the path. The p-values from this second approach are then corrected for multiple testing using the Bonferroni-Holm procedure. The second approach might be inappropriate, see also our discussion in Section 2.1 above pointing to the fact that the covariance test doesn't seem to test the hypotheses $H_{0,j}$; but for the sake of comparison (and practical use of the covariance test) we see no immediate other way to use the covariance test for constructing p-values for fixed hypotheses. For the $p > n$ situations, we use for all methods the variance estimator from the scaled Lasso (Sun and Zhang, 2012).

The results for $n = 100$ and $p = 80$ are reported in Table 1 and the results for $n = 100$ and $p = 200$ can be found in Table 2. In all settings, the de-sparsified Lasso method reliably controls the familywise error rate. In the $n > p$ setting, the covariance test has reasonable power at the cost of no control of the familywise error rate. In the $p > n$ setting, the covariance

test seems to be very conservative.

| $\text{FWER}_{\text{de-spars}}$ | $\text{TP}_{\text{de-spars}}$ | $\text{FWER}_{\text{cov}}$ | $\text{TP}_{\text{cov}}$ | $\text{FWER}_{\text{cov.pval}}$ | $\text{TP}_{\text{cov.pval}}$ |
|---|---|---|---|---|---|
| 0.042 | 2.626 | 0.072 | 1.304 | 0.020 | 0.736 |
| 0.056 | 7.104 | 0.124 | 2.884 | 0.064 | 3.770 |
| 0.064 | 9.116 | 0.284 | 5.992 | 0.210 | 7.556 |
| 0.064 | 9.478 | 0.426 | 8.394 | 0.298 | 9.324 |

TABLE 1

*(Empirical) familywise error rate (FWER) and average number of true positives (TP) for de-sparsified Lasso (*$\text{de}-\text{spars}$*) and both approaches of the covariance test (*cov *and* cov.pval*). The different rows correspond to coefficient size 0.5, 1, 2 and 4 (top to bottom). Sample size* $n = 100$ *and dimension* $p = 80$*.*

| $\text{FWER}_{\text{de-spars}}$ | $\text{TP}_{\text{de-spars}}$ | $\text{FWER}_{\text{cov}}$ | $\text{TP}_{\text{cov}}$ | $\text{FWER}_{\text{cov.pval}}$ | $\text{TP}_{\text{cov.pval}}$ |
|---|---|---|---|---|---|
| 0.030 | 1.320 | 0.012 | 0.416 | 0.002 | 0.120 |
| 0.046 | 3.304 | 0.010 | 0.632 | 0.004 | 0.274 |
| 0.052 | 4.934 | 0.018 | 0.956 | 0.006 | 0.860 |
| 0.060 | 5.594 | 0.032 | 1.550 | 0.018 | 1.884 |

TABLE 2

*(Empirical) familywise error rate (FWER) and average number of true positives (TP) for de-sparsified Lasso (*$\text{de}-\text{spars}$*) and both approaches of the covariance test (*cov *and* cov.pval*). The different rows correspond to coefficient size 0.5, 1, 2 and 4 (top to bottom). Sample size* $n = 100$ *and dimension* $p = 200$*.*

**6. Conclusions.** The authors present a novel and original idea of significance testing for "random hypotheses". In complex data scenarios, the strategy of considering "data-driven" hypotheses is certainly interesting, and the topic deserves further attention. The proposed solution to deal with such "random hypotheses" is based on strong beta-min or zonal assumptions, and this is somewhat unsatisfactory. The idea of taking the selection effect into account appears in other work for controlling the (Bayesian) false discovery rate (Benjamini and Yekutieli, 2005; Hwang and Zhao, 2013, cf.). We think that recent alternative approaches, as outlined in Section 5, are often more powerful and simpler to interpret when adopting the classical framework of (multiple) fixed hypotheses testing. It is an open question though whether the classical framework is the most appropriate tool for assigning "relevance" of single or groups of variables in complex or high-dimensional settings.

## References.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100:71–81.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid postselection inference. *Annals of Statistics*, 41:802–837.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242.

Bühlmann, P. and Mandozzi, J. (2013). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics*. Published online DOI 10.1007/s00180-013-0436-3.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag.

Chatterjee, A. and Lahiri, S. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics*, 41:1232–1259.

Ghosh, M., Reid, N., and Fraser, D. (2010). Ancillary statistics: a review. *Statistica Sinica*, 20:1309–1332.

Hwang, J. G. and Zhao, Z. (2013). Empirical Bayes confidence intervals for selected parameters in high-dimensional data. *Journal of the American Statistical Association*, 108:607–618.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. arXiv:1306.3171.

Leeb, H. and Pötscher, B. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. *Econometric Theory*, 19:100–142.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681.

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *J. of the American Statistical Association*, 106:1371–1382.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99:879–898.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv:1303.0518v2.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37:2178–2201.

Zhang, C.-H. and Zhang, S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society, Series B (to appear); arXiv:1110.2563*.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

E-MAIL: buhlmann@stat.math.ethz.ch E-MAIL: meier@stat.math.ethz.ch

E-MAIL: geer@stat.math.ethz.ch