Quantization Functionals and Regularized Principal Manifolds

Alex J. Smola, GMD [1] Sebastian Mika, GMD [2]
Bernhard Schölkopf, GMD [3]

NeuroCOLT2 Technical Report Series

NC2-TR-1998-028

September, 1998[4]

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

For more information see the NeuroCOLT website
http://www.neurocolt.com
or email neurocolt@neurocolt.com

[1]smola@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[2]mika@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[3]bs@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[4]Received 21-SEP-98

Abstract

Many settings of unsupervised learning can be viewed as quantization problems, namely of minimizing the expected quantization error subject to some restrictions. This has the advantage that tools known from the theory of (supervised) risk minimization like regularization can be readily applied to unsupervised settings. Moreover, one may show that this setting is very closely related to both, principal curves with a length constraint and the generative topographic map. Experimental results demonstrate the feasibility of the proposed method.

In a companion paper we show that uniform convergence bounds can be given for algorithms such as a modified variant of the principal curves problem.

1 Introduction

The problems in unsupervised learning are by far less precisely defined than in the supervised counterpart. Usually no explicit cost function exists with desired outputs or anything alike. Instead, one has to make assumptions on the data, with respect to which several questions may be asked.

One could look for properties of the data that can be extracted with high confidence. In other words, which feature extracting functions can be found among a given class with, say, unit variance and zero mean. Moreover, the properties should not change too much on unseen data. This leads to a *feature extracting* approach like (Kernel) Principal Component Analysis [9].

Another strategy is to look for properties that represent the data best. This means that one is looking for a *descriptive* model of the data, thus also a (possibly quite crude) model of the underlying probability distribution. Generative models like Principal Curves [5], the Generative Topographic Mapping [2], several linear Gaussian models [8], or also simple vector quantizers [1] are examples thereof. This is the kind of models we will study in this paper. We will show that many of the problems can be formalized in a quantization error setting. This will allow to use techniques from regularization theory. Moreover, we show in a companion paper [11] that one can use these results to give uniform convergence bounds.

2 The Quantization Error Functional

Denote $\mathcal{X}$ a vector space and $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}$ a dataset drawn iid from an underlying probability distribution $P(x)$. Moreover consider (compact) index sets $\mathcal{Z}$, maps $f : \mathcal{Z} \to \mathcal{X}$, and classes $\mathcal{F}$ of such maps (with $f \in \mathcal{F}$).
Here the map f is supposed to describe some basic properties of $P(x)$. In particular one seeks such f that the so–called quantization error

$$R[f] := \int_{\mathcal{X}} \min_{z \in \mathcal{Z}} \|x - f(z)\|^2 dP(x) \tag{1}$$

is minimized. Unfortunately, this is unsolvable, as P is in general unknown. Hence one replaces P by the empirical density $p(x) = \frac{1}{m}\sum_{i=1}^{m} \delta(x - x_i)$ and instead of (1) analyzes the empirical quantization error defined by

$$R_{\mathrm{emp}}[f] := \frac{1}{m}\sum_{i=1}^{m}\min_{z\in\mathcal{Z}}\|x_i - f(z)\|^2. \tag{2}$$

Many problems of unsupervised learning can be cast in the form of finding a minimizer of (1) or (2). Consider some practical examples.

Example 1 (Sample Mean) *Define $\mathcal{Z} := \{1\}$, $f : 1 \to f_1$ with $f_1 \in \mathcal{X}$, and $\mathcal{F}$ to be the set of all such functions. Then the minimum of*

$$R[f] := \int_{\mathcal{X}} \|x - f_1\|^2 dP(x) \tag{3}$$

denotes the variance of the data and the minimizers of the quantization functionals can be determined analytically by

$$\operatorname*{argmin}_{f\in\mathcal{F}} R[f] = \int_{\mathcal{X}} x dP(x) \text{ and } \operatorname*{argmin}_{f\in\mathcal{F}} R_{\mathrm{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} x_i. \tag{4}$$

This is the (empirical) sample mean. Via the law of large numbers it follows that both $R_{\mathrm{emp}}[f]$ and its minimizer converge to $R[f]$ and the corresponding minimizer.

Example 2 (k–Vectors Quantization) *Define $\mathcal{Z} := \{1, \ldots, k\}$, $f : i \to f_i$ with $f_i \in \mathcal{X}$, and $\mathcal{F}$ to be the set of all such functions. Then*

$$R[f] := \int_{\mathcal{X}} \min_{z\in\{1,\ldots,k\}} \|x - f_z\|^2 dP(x) \tag{5}$$

denotes the canonical distortion error of a vector quantizer. In practice one uses the k–means algorithm to find a set of vectors $\{f_1, \ldots, f_k\}$ minimizing the empirical quantization error. Also in this case, one can prove convergence properties of (the minimizer) of $R_{\mathrm{emp}}[f]$ to (the one of) $R[f]$ [1].

Instead of discrete quantization one can also consider a quantizer mapping the data onto a manifold of lower dimensionality than the input space. PCA can also be viewed in this way [5]. This is formalized in the following example:

Example 3 (Principal Components) *Define $\mathcal{Z} := [0, 1]$, $f : z \to f_0 + z \cdot f_1$ with $f_0, f_1 \in \mathcal{X}$, $\|f_1\| = 1$, and $\mathcal{F}$ to be the set of all such line segments. Then the minimizer of*

$$R[f] := \int_{\mathcal{X}} \min_{z\in[0,1]} \|x - f_0 - z \cdot f_1\|^2 dP(x) \tag{6}$$

yields a line segment parallel to the direction of largest variance in $P(x)$ [5].

Based on the properties of the current example, Hastie & Stuetzle [5] carried this idea further by also allowing other functions $f(z)$ than linear ones.

Example 4 (Principal Curves) *Denote $\mathcal{Z} := [0,1]^D$ (with $D > 1$ for principal surfaces), $f : z \to f(z)$ with $f \in \mathcal{F} \subseteq \mathcal{C}^0[0,1]^D$, i.e. the class of continuous curves, possibly with a further restriction of $\mathcal{F}$. The minimizer of*

$$R[f] := \int_{\mathcal{X}} \min_{z \in [0,1]^D} \|x - f(z)\|^2 dP(x) \tag{7}$$

is not well defined, unless $\mathcal{F}$ is a compact set. Moreover, even the minimizer of $R_{\mathrm{emp}}[f]$ is not well defined either, in general. In fact, it is an ill posed problem in the sense of Arsenin and Tikhonow [12]. Until recently [6], no convergence properties of $R_{\mathrm{emp}}[f]$ to $R[f]$ could be stated.

Despite the problems mentioned above, an algorithm to minimize $R_{\mathrm{emp}}[f]$, was devised by [5]. It proceeds as follows: after initialization to the principal components, the projections of the data-points onto the curve are estimated, the curve based on that is re–estimated and smoothed by kernel smoothers or similar techniques. This is iterated until a fixed point has been reached.

Kegl et al. [6, 7] modified the original "principal–curves" algorithm slightly, in order to prove bounds on $R[f]$ wrt. $R_{\mathrm{emp}}[f]$ and to show that the resulting estimate is well defined. In particular the changes imply a restriction of $\mathcal{F}$ to polygonal lines with a fixed number of knots and, most importantly, *fixed* length L.

Instead of a length constraint we now consider smoothness constraints on the estimated curve $f(x)$. This is done via a regularization operator.

3 Invariant Regularizers

As a first step we will show that the class of admissible operators can be restricted to scalar ones, provided some basic assumption about scaling behavior and permutation symmetry are imposed.

Proposition 1 (Homogeneous Invariant Regularization) *Any regularization term $Q[f]$ that is both, homogeneous quadratic and invariant under an irreducible orthogonal representation ρ of the group $\mathcal{G}$ on $\mathcal{X}$, i.e. satisfies*

$$Q[f] \geq 0 \text{ for all } f \in \mathcal{F} \tag{8}$$

$$Q[af] = a^2 Q[f] \text{ for all scalars } a \tag{9}$$

$$Q[\rho(g)f] = Q[f] \text{ for all } \rho(g) \in \mathcal{G} \tag{10}$$

is of the form

$$Q[f] = \langle Pf, Pf \rangle \text{ where } P \text{ is a "scalar" operator.} \tag{11}$$

Proof It follows directly from (9) and Euler's "homogeneity property", that $Q[f]$ has to be a quadratic form, thus $Q[f] = \langle f, Mf \rangle$ for some operator M. Moreover M can be written as P^*P as it has to be a positive operator (cf. (8)). Finally from

$$\langle Pf, Pf \rangle = \langle P\rho(g)f, P\rho(g)f \rangle \tag{12}$$

and the polarization equation it follows that $P^*P\rho(g) = \rho(g)P^*P$ has to hold for any $\rho(g) \in \mathcal{G}$. Thus, by virtue of Schur's lemma (cf. e.g. [4]), it follows that P^*P only may be a scalar operator. Then, without loss of generality, also P may be assumed to be scalar. ∎

A consequence of the proposition above is that there exists no "vector valued" regularization operator satisfying the invariance conditions. Hence it is useless to look for other operators P in the presence of a sufficiently strong invariance. A practical application of proposition 1 is the following corollary.

Corollary 2 (Permutation and Rotation Symmetries) *Under the assumptions of proposition 1 both, the canonical representation of the permutation group in a finite dimensional vector space $\mathcal{X}$ and the group of orthogonal transformations on $\mathcal{X}$ enforce scalar operators P.*

This follows immediately from the fact that these groups are unitary and irreducible on $\mathcal{X}$ by construction. Thus in the following we will only consider scalar operators P.

4 A Regularized Quantization Functional

In the following a modification to minimizing the empirical quantization functional is proposed, which will lead to an algorithm that is more amenable to implementation. Moreover, uniform convergence bounds can be obtained for smooth curves, independently of the number of nodes/grid-points [11]. For this purpose, a regularized version of the empirical quantization functional is needed.

$$R_{\text{reg}}[f] := R_{\text{emp}}[f] + \frac{\lambda}{2}\|Pf\|^2 = \sum_{i=1}^{m} \min_{z \in \mathcal{Z}} \|x_i - f(z)\|^2 + \frac{\lambda}{2}\|Pf\|^2. \tag{13}$$

Here P is a *scalar* regularization operator in the sense of Arsenin and Tikhonov, penalizing unsmooth functions f (see [10] for details). In the present case this is a useful assumption, since all curves, which can be transformed into each other by rotations, should be penalized equally.
Using the results of [10] regarding the connection between regularization operators and kernels it appears suitable to choose a kernel expansion of f matching the regularization operator P, i.e. $\langle Pk(x_i, \cdot), Pk(x_j, \cdot)\rangle = k(x_i, x_j)$. Finally assume $P^*Pf_0 = 0$, i.e. constant functions are not regularized. Hence one gets

$$f(z) = f_0 + \sum_{i=1}^{M} \alpha_i k(z_i, z) \text{ with } z_i \in \mathcal{Z},\ \alpha_i \in \mathcal{X},\ \text{and } k : \mathcal{Z}^2 \to \mathbb{R}. \tag{14}$$

for some previously chosen nodes $z_1, \ldots, z_M$ (of which one takes as many as one may afford in terms of computational cost). Consequently the regularization term can be written as

$$\|Pf\|^2 = \sum_{i,j=1}^{M} \langle \alpha_i, \alpha_j \rangle k(z_i, z_j). \tag{15}$$

What remains to do is to find an algorithm that finds a minimizer of R_{emp}. This is achieved by an EM type strategy. In the following we will assume the data to be centered and therefore drop the term f_0. This greatly simplifies the notation (while the modification to take f_0 into account is straightforward).

4.1 An Algorithm for minimizing $R_{\text{reg}}[f]$

No re–interpretation of the regularized quantization error as some likelihood (with a suitable prior) of a class of generative models is done. Instead, the techniques of EM algorithms [3] are adapted to solve

$$\min_{\substack{\{\alpha_1,\ldots,\alpha_M\}\subset\mathfrak{X}\\ \{\zeta_1,\ldots,\zeta_m\}\subset\mathcal{Z}}} \left[\sum_{i=1}^{m}\left\|x_i - \sum_{j=1}^{M}\alpha_j k(\zeta_i, z_j)\right\|^2 + \frac{\lambda}{2}\sum_{i,j=1}^{M}\langle\alpha_i,\alpha_j\rangle k(z_i,z_j)\right] \tag{16}$$

likewise in an iterative fashion. For this purpose one iterates over minimizing (16) with respect to $\{\zeta_1,\ldots,\zeta_m\}$, equivalent to the projection step, and $\{\alpha_1,\ldots,\alpha_M\}$, which corresponds to the expectation step. This is repeated until convergence, in practice until the regularized quantization functional does not decrease significantly any further. One obtains:

Projection For each $i \in \{1,\ldots,m\}$ choose ζ_i such that

$$\zeta_i := \operatorname*{argmin}_{\zeta\in\mathcal{Z}} \|f(\zeta) - x_i\|^2. \tag{17}$$

Clearly, for fixed α_i, the so chosen ζ_i minimize the term in (16), which in turn is equal to $R_{\text{reg}}[f]$ for given α_i and X.

Adaptation Now the parameters ζ_i are fixed and α_i is adapted such that $R_{\text{reg}}[f]$ decreases further. For fixed ζ_i differentiation of (16) with respect to α_i yields

$$\left(\frac{\lambda}{2}K_z + K_\zeta^\top K_\zeta\right)\alpha = K_\zeta^\top X \tag{18}$$

where $(K_z)_{ij} := k(z_i, z_j)$ is an $M\times M$ matrix and $(K_\zeta)_{ij} := k(\zeta_i, z_j)$ is $m\times M$. Moreover, with slight abuse of notation, α, and X denote the *matrix* of all parameters, and samples, respectively. The term in (16) keeps on decreasing until the algorithm converges to a (local) minimum. What remains is to find good starting values.

Initialization If not dealing, as assumed, with centered data, set f_0 to the sample mean, i.e. $f_0 = \frac{1}{m}\sum_{i=1}^{m} x_i$. Moreover, choose the coefficients α_i such that f approximately points into the directions of the first D principal components given by the matrix $E := (e_1,\ldots,e_D)$. This is done as follows, analogously to the initialization in the generative topographic map [2, eq. (20)].

$$\min_{\{\alpha_1,\ldots,\alpha_M\}\subset\mathfrak{X}} \left[\sum_{i=1}^{M}\left\|E(z_i - z_0) - \sum_{j=1}^{M}\alpha_j k(z_i, z_j)\right\|^2 + \frac{\lambda}{2}\sum_{i,j=1}^{M}\langle\alpha_i,\alpha_j\rangle k(z_i,z_j)\right]. \tag{19}$$

Thus α is determined as the solution of $\left(\frac{\lambda}{2}\mathbf{1} + K_z\right)\alpha = E(Z - Z_0)$ where Z denoted the matrix of z_i, z_0 the mean of z_i and Z_0 the corresponding matrix.

The derivation of this algorithm was quite "ad hoc", however, one can show that there exist similar precursors in the literature. First it is shown that minimizing (13) is equivalent to minimizing the quantization error subject to a length constraint on the estimated curve.

4.2 Regularizers for Length Constraints

By choosing $P := \partial_z$, i.e. the differentiation operator, $\|Pf\|^2$ becomes an integral over the squared "speed" of the curve. Reparameterizing f to constant speed leaves the empirical quantization error unchanged, whereas the regularization term is minimized. This can be seen as follows: by construction $\int_{[0,1]} \|\partial_z f(z)\| dz$ does not depend on the (re)parameterization. The variance, however, is minimal for a constant function, hence $\|\partial_z f(z)\|$ has to be constant over interval $[0, 1]$. Thus $\|Pf\|^2$ equals the squared length L^2 of the curve at the optimal solution.

One can show that minimizing the empirical quantization error plus a regularizer is equivalent to minimizing the empirical quantization error for a fixed value of the regularization term (for λ adjusted suitably). Hence the proposed algorithm is equivalent to finding the optimal curve subject to a length constraint, i.e. it is equivalent to the algorithm proposed by [6].[1]

4.3 The Connection to the GTM

Just considering the basic algorithm of the GTM (without the Bayesian framework), one can observe that it minimizes a rather similar quantity to $R_{\mathrm{reg}}[f]$. It differs in its choice of $\mathcal{Z}$, which is chosen to be a grid, identical with the points z_i in our setting, and the different regularizer (called Gaussian prior in that case) which is of ℓ_2 type. In other words instead of using $\|Pf\|^2$ Bishop et al. [2] choose $\sum_i \|\alpha_i\|^2$ as a regularizer. Finally in the GTM several ζ_i may take on "responsibility" for having generated a data-point x_i (this follows naturally from the generative model setting in the latter case).

Note that unlike in the GTM (cf. [2, sec. 2.3]) the number of nodes (for the kernel expansion) is not a critical parameter. This is due to the fact that there is a *coupling* between the single centers of the basis functions $k(z_i, z_j)$ via the regularization operator. If needed, one could also see the proposed algorithm in a Gaussian Process context (see [13]) — the data X then should be interpreted as created by a homogeneous process mapping from $\mathcal{Z}$ to $\mathcal{X}$.

Finally the use of periodical kernels (cf. [10]) allows one to model circular structures in $\mathcal{X}$. After solving the algorithmic issue one has to come up with good uniform convergence bounds [11].

---

[1] The reasoning is slightly incorrect — f *cannot* be completely reparameterized to constant speed, as it is an expansion in terms of a *finite* number of nodes. However the basic properties still hold.

5 Experiments

In order to show that the basic idea of the proposed algorithm is sound, we ran several toy experiments (cf. figure 1). We generated different data sets in 2 and 3 dimensions from 1 or 2 dimensional parameterizations. Then we applied our algorithm using the prior knowledge about the original parameterization dimension of the data set in choosing the latent variable space to have the appropriate size. For almost any parameter setting (λ, M, and width of basis functions) we obtained reasonable results.

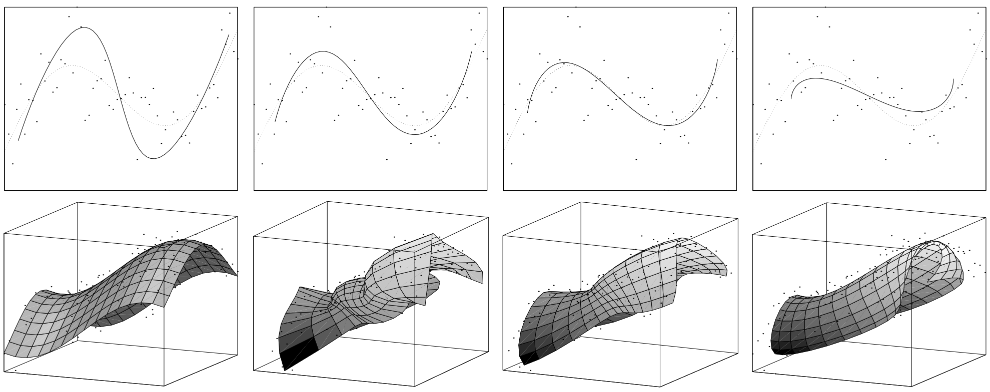

Figure 1 Upper 4 images. We generated a dataset (small dots) by adding noise to a distribution indicated by the dotted line. The resulting manifold generated by our approach is given by the solid line (over a parameter range of $\mathcal{Z} = [-1, 1]$). From left to right we used different values for the regularization parameter $\lambda = 0.1, 0.5, 1, 4$. The width and number of basis function was constant 1, and 10 respectively. Lower 4 images. Here we generated a dataset by sampling (with noise) from a distribution depicted in the left most image (small dots are the sampled data). The remaining three images show the manifold yielded by our approach over the parameter space $\mathcal{Z} = [-1, 1]^2$ for $\lambda = 0.001, 0.1, 1$. The width and number of basis functions was constant again (1 and 36).

We found, that for a suitable choice of the regularization factor λ a very close match to the original distribution can be achieved. The number and width of the basis functions had of course an effect on the solution, too. But their influence on the basic characteristics is quite small.

Finally, figure 2 shows the convergence properties of the algorithm. One can clearly observe that the overall regularized quantization error decreases for each step, while both the regularization term and the quantization error term are free to vary. This experimentally shows that the algorithm finds a (local) minimum of $R_{\text{quant}}[f]$.

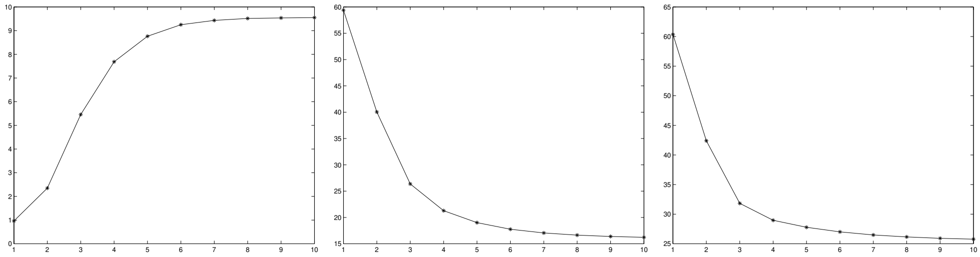

Figure 2 Left: regularization term, middle: empirical quantization error, right: regularized quantization error vs. number of iterations.

6 Discussion

We proposed a framework for unsupervised learning that can draw on the techniques available in minimization of risk functionals in supervised learning. This yielded an algorithm suitable to deal with principal manifolds. The expansion in terms of kernel functions and the treatment by regularization operators made it easier to decouple the algorithmic part (of finding a suitable manifold) from the part of specifying a class of manifolds with desirable properties. In particular, our algorithm does not crucially depend on the number of nodes used. Furthermore the regularization operator treatment makes it easier to obtain uniform convergence results (cf. [11]).

The current approach builds a bridge between algorithms such as Principal Curves/Surfaces and the Generative Topographic Map, two methods that did not seem too closely related after all. Using the proposed techniques, it should be quite straightforward, to modify the GTM in a way to take a Gaussian process prior into account.

It is our hope that, building on the current results, it will be possible to use the framework of regularized risk functionals for capacity control in an effective way also in unsupervised learning.

Acknowledgments This work was supported in part by grants of the ARC and the DFG (Ja 379/71 and Ja 379/51). Moreover we thank Adam Krzyzak for helpful comments and discussions.

References

[1] P. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. Technical report, Australian National University, RSISE, 1997. extended abstract in Eurocolt'97.

[2] C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: The generative topographic mapping. Technical Report NCRG/96/015, Neural Computing Research Group, Aston University, UK, 1996. to appear in Neural Computation.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc. B*, 39:1–38, 1977.

[4] M. Hamermesh. *Group theory and its applications to physical problems.* Addison Wesley, Reading, MA, 2 edition, 1962. Reprint by Dover, New York, NY.

[5] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[6] B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. submitted, URL=http://magenta.mast.queensu.ca/˜linder/psfiles/KeKrLiZe97.ps.

[7] B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. A polygonal line algorithm for constructing principal curves. In *NIPS '98*. MIT Press, 1999. forthcoming.

[8] S. Roweis and Z. Ghabramani. A unifying review of linear gaussian models. *Neural Computation*, 8, 1998. to appear.

[9] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299 – 1319, 1998.

[10] A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[11] A.J. Smola, R.C. Williamson, and B. Schölkopf. Generalization bounds and learning rates for regularized principal manifolds. Technical Report NC-TR-98-027, Royal Holloway College, University of London, UK, 1998. submitted to EUROCOLT 99.

[12] A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill–Posed Problems.* Winston, Washington, DC, 1977.

[13] C.K.I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. *Learning and Inference in Graphical Models*, 1998. also Technical Report at NCRG/97/012.