Dictionary-Free MRI PERK: Parameter Estimation via Regression with Kernels

Gopal Nataraj, *Student Member, IEEE*, Jon-Fredrik Nielsen, Clayton Scott, *Member, IEEE*, and Jeffrey A. Fessler, *Fellow, IEEE*

***Abstract*— This paper introduces a fast, general method for dictionary-free parameter estimation in quantitative magnetic resonance imaging (QMRI) parameter estimation via regression with kernels (PERK). PERK first uses prior distributions and the nonlinear MR signal model to simulate many parameter-measurement pairs. Inspired by machine learning, PERK then takes these parameter-measurement pairs as labeled training points and learns from them a nonlinear regression function using kernel functions and convex optimization. PERK admits a simple implementation as per-voxel nonlinear lifting of MRI measurements followed by linear minimum mean-squared error regression. We demonstrate PERK for T_1, T_2 estimation, a well-studied application where it is simple to compare PERK estimates against dictionary-based grid search estimates and iterative optimization estimates. Numerical simulations as well as single-slice phantom and *in vivo* experiments demonstrate that PERK and other tested methods produce comparable T_1, T_2 estimates in white and gray matter, but PERK is consistently at least 140× faster. This acceleration factor may increase by several orders of magnitude for full-volume QMRI estimation problems involving more latent parameters per voxel.**

***Index Terms*— Nonlinear regression, parameter mapping, magnetic resonance imaging, machine learning, kernels.**

I. Introduction

IN QUANTITATIVE magnetic resonance imaging (QMRI), one seeks to estimate latent parameter images from suitably informative data. Since MR acquisitions are tunably sensitive to many physical processes (*e.g.*, relaxation [1], diffusion [2], and chemical exchange [3]), MRI parameter estimation is important for many QMRI applications (*e.g.*, relaxometry [4], diffusion tensor imaging [5], and multi-compartmental imaging [6]). Motivated by widespread applications, this manuscript introduces a general method for fast MRI parameter estimation.

Manuscript received February 20, 2018; accepted March 13, 2018. Date of publication March 20, 2018; date of current version August 30, 2018. This work was supported in part by the National Institutes of Health under Grant P01 CA87634 and in part by the University of Michigan through an MCubed Seed Grant and a Predoctoral Fellowship. *(Corresponding author: Gopal Nataraj.)*

G. Nataraj, C. Scott, and J. A. Fessler are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: gnataraj@umich.edu; clayscot@umich.edu; fessler@umich.edu).

J.-F. Nielsen is with the Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109, USA (e-mail: jfnielse@umich.edu).

This paper has supplementary material provided by the authors available for download at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2018.2817547

A common MRI parameter estimation strategy involves minimizing a cost function related to a statistical likelihood function. Because MR signal models are typically nonlinear functions of the underlying latent parameters, such likelihood-based estimation usually requires non-convex optimization. To seek good solutions, many works (*e.g.*, [7]–[21]) approach estimation with algorithms that employ exhaustive grid search, which requires either storing or computing on-the-fly a "dictionary" of signal vectors. These works estimate a small number (2-3) of nonlinear latent parameters, so grid search is practical. However, for moderate or large sized problems, the required number of dictionary elements renders grid search undesirable or even intractable, unless one imposes artificially restrictive latent parameter constraints. Though several recent works [13], [17], [20], [21] focus on reducing dictionary storage requirements, all of these methods ultimately rely on some form of dictionary-based grid search.

There are numerous QMRI applications that could benefit from an alternative parameter estimation method that scales well with the number of latent parameters. For example, vector (*e.g.*, flow [22]) and tensor (*e.g.*, diffusivity [5] or conductivity [23]) field mapping techniques require estimation of at minimum 4 and 7 latent parameters per voxel, respectively. Phase-based longitudinal [24] or transverse [25], [26] field mapping could avoid noise-amplifying algebraic manipulations on reconstructed image data that are conventionally used to reduce signal dependencies on nuisance latent parameters. Compartmental fraction mapping [6], [27] from steady-state pulse sequences requires estimation of at least 7 [28] and as many as 10 [29] latent parameters per voxel. In these and other applications, greater estimation accuracy requires more complete signal models that involve more latent parameters, increasing the need for scalable estimation methods.

The fundamental challenge of scalable MRI parameter estimation stems from MR signal model nonlinearity: standard linear estimators would be scalable but inaccurate. One natural solution strategy involves nonlinearly preprocessing reconstructed images such that the transformed images are at least approximately linear in the latent parameters. As an example, for simple T_2 estimation from measurements at multiple echo times, one could apply linear regression to the logarithm of the measurements (§S.I[1] builds further intuition using this simple application). However, such simple

[1]Supplementary material is available in the /media tab on IEEEXplore.

transformations are generally not evident for more complicated signal models. Without such problem-specific insight, sufficiently rich nonlinear transformations could dramatically increase problem dimensionality, hindering scalability. Fortunately, a celebrated result in approximation theory [30] showed that simple transformations involving *reproducing kernel* functions [31] can represent nonlinear estimators whose evaluation need not directly scale in computation with the (possibly very high) dimension of the associated transformed data. These kernel methods later found popularity in machine learning (initially for classification [32] and quickly thereafter for other applications, *e.g.*, regression [33]) because they provided simple, scalable nonlinear extensions to fast linear algorithms.

The general idea of using linearization to simplify a nonlinear estimation problem has been used before in QMRI. For example, orthogonal transforms have been used to linearly represent exponential [34] and extended phase graph [35] models for T_2 estimation. An unscented Kalman filter has been used to linearly represent nonlinear models for general multiple-parameter e stimation up to third-order accuracy [36]. Whereas these prior works largely focus on parameter estimation accuracy gains in under-sampled acquisitions, this paper focuses on acceleration for general per-voxel MRI parameter estimation from reconstructed images.

This paper introduces[2] a fast, dictionary-free method for MRI parameter estimation via regression with kernels (PERK). PERK first simulates many instances of latent parameter inputs and measurement outputs using prior distributions and a general nonlinear MR signal model. PERK takes such input-output pairs as simulated *training points* and then *learns* (using an appropriate nonlinear kernel function) a nonlinear *regression function* from the training points. PERK may scale considerably better with the number of latent parameters than likelihood-based estimation via grid search.

The remainder of this manuscript is organized as follows. §II reviews pertinent background information about kernels. §III formulates a function optimization problem for MRI parameter estimation and efficiently solves this problem using kernels. §IV studies bias and covariance of the resulting PERK estimator. §V addresses practical implementation issues such as computational complexity and model selection. §VI demonstrates PERK in numerical simulations as well as phantom and *in vivo* experiments. §VII discusses advantages, challenges, and extensions. §VIII summarizes key contributions.

II. Preliminaries

This brief section reviews relevant definitions and facts about kernels. A (real-valued) *kernel* $k : \mathbb{P}^2 \mapsto \mathbb{R}$ is a function that describes a measure of similarity between two pattern vectors $\mathbf{p}, \mathbf{p}' \in \mathbb{P}$. The matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ associated with kernel k and $N \in \mathbb{N}$ patterns $\mathbf{p}_1, \ldots, \mathbf{p}_N \in \mathbb{P}$ consists of entries $k(\mathbf{p}_n, \mathbf{p}_{n'})$ for $n, n' \in \{1, \ldots, N\}$. A *positive definite kernel* is a kernel for which $\mathbf{K}$ is positive semidefinite (PSD) for any finite set of pattern vectors, in which case $\mathbf{K}$ is a *Gram matrix*. A *symmetric kernel* satisfies $k(\mathbf{p}, \mathbf{p}') = k(\mathbf{p}', \mathbf{p})$ $\forall \mathbf{p}, \mathbf{p}' \in \mathbb{P}$. We hereafter restrict attention to symmetric, positive definite (SPD) kernels.

An SPD kernel $k : \mathbb{P}^2 \mapsto \mathbb{R}$ defines an inner product in a particular Hilbert function space $\bar{\mathbb{H}}$ that we briefly describe here because it characterizes the class of candidate regression functions over which PERK operates. To envision $\bar{\mathbb{H}}$, first define a kernel's associated *(canonical) feature map* $\mathbf{z} : \mathbb{P} \mapsto \mathbb{R}^{\mathbb{P}}$ that assigns each $\mathbf{p} \in \mathbb{P}$ to a *(canonical) feature* $k(\cdot, \mathbf{p}) \in \mathbb{R}^{\mathbb{P}}$. Then $\bar{\mathbb{H}}$ is a completion of the space $\mathbb{H} := \left\{ \sum_{n=1}^{N} a_n k(\cdot, \mathbf{p}_n) \right\}$ spanned by point evaluations of the feature map, where $N \in \mathbb{N}$, $a_1, \ldots, a_N \in \mathbb{R}$, and $\mathbf{p}_1, \ldots, \mathbf{p}_N \in \mathbb{P}$ are arbitrary. Let $\langle \cdot, \cdot \rangle : \bar{\mathbb{H}}^2 \mapsto \mathbb{R}$ denote the inner product on $\bar{\mathbb{H}}$. Then for any $h, h' \in \mathbb{H}$ that have finite-dimensional canonical representations $h := \sum_{n=1}^{N} a_n k(\cdot, \mathbf{p}_n)$ and $h' := \sum_{n'=1}^{N} b_{n'} k(\cdot, \mathbf{p}_{n'})$, the assignment

$$\langle h, h' \rangle_{\bar{\mathbb{H}}} = \sum_{n=1}^{N} \sum_{n'=1}^{N} a_n b_{n'} k(\mathbf{p}_{n'}, \mathbf{p}_n) \tag{1}$$

is consistent with the inner product on $\bar{\mathbb{H}}$. This inner product exhibits $\forall h \in \bar{\mathbb{H}}, \mathbf{p} \in \mathbb{P}$ an interesting *reproducing property*

$$\langle h, k(\cdot, \mathbf{p}) \rangle_{\bar{\mathbb{H}}} = h(\mathbf{p}) \tag{2}$$

that can be seen to directly follow from (1) for $h \in \mathbb{H}$.

A *reproducing kernel* (RK) is a kernel that satisfies (2) for some real-valued Hilbert space $\bar{\mathbb{H}}$. A kernel is reproducing if and only if it is SPD. There is a bijection between RK k and $\bar{\mathbb{H}}$, and so $\bar{\mathbb{H}}$ is often called the *reproducing kernel Hilbert space* (RKHS) uniquely associated with RK k. This bijection is critical to practical function optimization over an RKHS in that it translates inner products in a (usually high-dimensional) RKHS $\bar{\mathbb{H}}$ into equivalent kernel operations in the (lower-dimensional) pattern vector space $\mathbb{P}$. The following sections exploit the bijection between an RKHS and its associated RK.

III. A Function Optimization Problem and Kernel Solution for MRI Parameter Estimation

After image reconstruction, many QMRI acquisitions produce at each voxel position a sequence of noisy measurements $\mathbf{y} \in \mathbb{C}^D$, modeled as

$$\mathbf{y} = \mathbf{s}(\mathbf{x}, \boldsymbol{\nu}) + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^L$ denotes L *latent* parameters; $\boldsymbol{\nu} \in \mathbb{R}^K$ denotes K *known* parameters; $\mathbf{s} : \mathbb{R}^L \times \mathbb{R}^K \mapsto \mathbb{C}^D$ models D noiseless continuous signal functions; and $\boldsymbol{\epsilon} \sim \mathbb{C}\mathcal{N}(\mathbf{0}_D, \boldsymbol{\Sigma})$ is complex Gaussian noise with zero mean $\mathbf{0}_D \in \mathbb{R}^D$ and known covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$. (As a concrete example, for T_2 estimation from single spin echo measurements, $\mathbf{x}$ could collect spin density and T_2; $\boldsymbol{\nu}$ could collect known longitudinal and transverse field inhomogeneities; and $\mathbf{y}$ could collect measurements at D echo times.) We seek to estimate on a per-voxel basis each latent parameter $\mathbf{x}$ from corresponding measurement $\mathbf{y}$ and known parameter $\boldsymbol{\nu}$.

To develop an estimator $\widehat{\mathbf{x}}$, we simulate many instances of forward model (3) and use kernels to estimate a nonlinear inverse function. We sample part of $\mathbb{R}^L \times \mathbb{R}^K \times \mathbb{C}^D$

[2]This manuscript substantially extends [37], our conference paper that recently introduced kernel-based MRI parameter estimation. Though kernels have been used in MRI reconstruction (*e.g.*, [38], [39]) and MRI analysis (*e.g.*, [40], [41]), kernels had not to our knowledge been used prior to [37] for MRI parameter estimation.

and evaluate (3) N times to produce sets of parameter and noise realizations $\{(\mathbf{x}_1, \boldsymbol{\nu}_1, \boldsymbol{\epsilon}_1), \ldots, (\mathbf{x}_N, \boldsymbol{\nu}_N, \boldsymbol{\epsilon}_N)\}$ and corresponding measurements $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. We seek a function $\widehat{\mathbf{h}} : \mathbb{R}^P \mapsto \mathbb{R}^L$ and an offset $\widehat{\mathbf{b}} \in \mathbb{R}^L$ that together map each pure-real[3] regressor $\mathbf{p}_n := [|\mathbf{y}_n|^\mathsf{T}, \boldsymbol{\nu}_n^\mathsf{T}]^\mathsf{T}$ to an estimate $\widehat{\mathbf{x}}(\mathbf{p}_n) := \widehat{\mathbf{h}}(\mathbf{p}_n) + \widehat{\mathbf{b}}$ that is "close" to corresponding regressand $\mathbf{x}_n$, where $P := D + K$, $n \in \{1, \ldots, N\}$, and $(\cdot)^\mathsf{T}$ denotes vector transpose. For any finite N, there are infinitely many candidate estimators that are consistent with training points in this manner. We use function regularization to choose one estimator that smoothly interpolates between training points:

$$\big(\widehat{\mathbf{h}}, \widehat{\mathbf{b}}\big) \in \arg \min_{\substack{\mathbf{h} \in \bar{\mathbb{H}}^L \\ \mathbf{b} \in \mathbb{R}^L}} \Psi\Big(\mathbf{h}, \mathbf{b}; \{(\mathbf{x}_n, \mathbf{p}_n)\}_1^N\Big), \tag{4}$$

where

$$\Psi(\ldots) = \sum_{l=1}^{L} \Psi_l\Big(h_l, b_l; \big\{(x_{l,n}, \mathbf{p}_n)\big\}_1^N\Big); \tag{5}$$

$$\Psi_l(\ldots) = \rho_l \|h_l\|_{\bar{\mathbb{H}}}^2 + \frac{1}{N} \sum_{n=1}^{N} \big(h_l(\mathbf{p}_n) + b_l - x_{l,n}\big)^2. \tag{6}$$

Here, each $h_l : \mathbb{R}^P \mapsto \mathbb{R}$ is a scalar function that maps to the lth component of the output of $\mathbf{h}$; each $b_l, x_{l,n} \in \mathbb{R}$ are scalar components of $\mathbf{b}, \mathbf{x}_n$; $\bar{\mathbb{H}}$ is an RKHS whose norm $\|\cdot\|_{\bar{\mathbb{H}}}$ is induced by inner product $\langle \cdot, \cdot \rangle_{\bar{\mathbb{H}}} : \bar{\mathbb{H}}^2 \mapsto \mathbb{R}$; and each ρ_l controls for regularity in h_l.

Since (5) is separable in the components of $\mathbf{h}$ and $\mathbf{b}$, it suffices to consider optimizing each (h_l, b_l) by separately minimizing (6) for each $l \in \{1, \ldots, L\}$. Remarkably, a generalization of the Representer Theorem [42], restated as is relevant here for completeness, reduces minimizing (6) to a finite-dimensional optimization problem.

Theorem 1 (Generalized Representer, [42]): Define $k : \mathbb{R}^Q \times \mathbb{R}^Q \mapsto \mathbb{R}$ *to be the SPD kernel associated with RKHS* $\bar{\mathbb{H}}$*, such that reproducing property* $h_l(\mathbf{p}) = \langle h_l, k(\cdot, \mathbf{p}) \rangle_{\bar{\mathbb{H}}}$ *holds for all* $h_l \in \bar{\mathbb{H}}$ *and* $\mathbf{p} \in \mathbb{R}^Q$*. Then any minimizer* $(\widehat{h}_l, \widehat{b}_l)$ *of* (6) *over* $\bar{\mathbb{H}} \times \mathbb{R}$ *admits a representation for* $\widehat{h}_l$ *of the form*

$$\widehat{h}_l(\cdot) \equiv \sum_{n=1}^{N} a_{l,n} k(\cdot, \mathbf{p}_n), \tag{7}$$

where each $a_{l,n} \in \mathbb{R}$ *for* $n \in \{1, \ldots, N\}$.

Theorem 1 ensures that any solution to the component-wise $(N+1)$-dimensional problem

$$(\widehat{\mathbf{a}}_l, \widehat{b}_l) \in \arg \min_{\substack{\mathbf{a}_l \in \mathbb{R}^N \\ b_l \in \mathbb{R}}} \rho_l \Big\| \sum_{n'=1}^{N} a_{l,n'} k(\cdot, \mathbf{p}_{n'}) \Big\|_{\bar{\mathbb{H}}}^2 + \frac{1}{N} \sum_{n=1}^{N} \Bigg(\sum_{n'=1}^{N} a_{l,n'} k(\mathbf{p}_n, \mathbf{p}_{n'}) + b_l - x_{l,n} \Bigg)^2 \tag{8}$$

corresponds via (7) to a minimizer of (6) over $\bar{\mathbb{H}} \times \mathbb{R}$, where $\mathbf{a}_l := [a_{l,1}, \ldots, a_{l,N}]^\mathsf{T}$. Fortunately, a solution of (8) exists uniquely for $\rho_l > 0$ and can be expressed as

$$\widehat{\mathbf{a}}_l = (\mathbf{MKM} + N\rho_l \mathbf{I}_N)^{-1} \mathbf{M} \boldsymbol{x}_l; \tag{9}$$

$$\widehat{b}_l = \frac{1}{N} \mathbf{1}_N^\mathsf{T} (\boldsymbol{x}_l - \mathbf{K}\widehat{\mathbf{a}}_l), \tag{10}$$

where $\mathbf{K} \in \mathbb{R}^{N \times N}$ is the Gram matrix consisting of entries $k(\mathbf{p}_n, \mathbf{p}_{n'})$ for $n, n' \in \{1, \ldots, N\}$; $\mathbf{M} := \mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\mathsf{T} \in \mathbb{R}^{N \times N}$ is a de-meaning operator; $\boldsymbol{x}_l := [x_{l,1}, \ldots, x_{l,N}]^\mathsf{T}$; $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ is the identity matrix; and $\mathbf{1}_N \in \mathbb{R}^N$ is a vector of ones. Substituting (9) into (7) yields an expression for the lth entry $\widehat{x}_l$ of MRI parameter estimator $\widehat{\mathbf{x}}$:

$$\widehat{x}_l(\cdot) \leftarrow \boldsymbol{x}_l^\mathsf{T} \big(\tfrac{1}{N} \mathbf{1}_N + \mathbf{M}(\mathbf{MKM} + N\rho_l \mathbf{I}_N)^{-1} \mathbf{k}(\cdot) \big), \tag{11}$$

where $\mathbf{k}(\cdot) := \big[k(\cdot, \mathbf{p}_1), \ldots, k(\cdot, \mathbf{p}_N)\big]^\mathsf{T} - \frac{1}{N} \mathbf{K} \mathbf{1}_N : \mathbb{R}^Q \mapsto \mathbb{R}^N$ is a kernel embedding operator.

When $\rho_l > 0 \ \forall l \in \{1, \ldots, L\}$, estimator $\widehat{\mathbf{x}}(\cdot)$ with entries (11) minimizes (5) over $\bar{\mathbb{H}}^L \times \mathbb{R}^L$. However, the utility of $\widehat{\mathbf{x}}(\cdot)$ depends on the choice of kernel k, which induces a choice on the RKHS $\bar{\mathbb{H}}$ and thus the function space $\bar{\mathbb{H}}^L \times \mathbb{R}^L$ over which (4) optimizes. For example, if k was selected as the canonical dot product $k(\mathbf{p}, \mathbf{p}') \leftarrow \langle \mathbf{p}, \mathbf{p}' \rangle_{\mathbb{R}^Q} := \mathbf{p}^\mathsf{T} \mathbf{p}'$ (for which RKHS $\bar{\mathbb{H}} \leftarrow \mathbb{R}^Q$), then (11) would reduce to affine ridge regression [43] which is optimal over $\mathbb{R}^Q \times \mathbb{R}$ but is unlikely to be useful when signal model $\mathbf{s}$ is nonlinear in $\mathbf{x}$. Since we expect a useful estimate $\widehat{\mathbf{x}}(\mathbf{p})$ to depend nonlinearly (but smoothly) on $\mathbf{p}$ in general, we instead use an SPD kernel k that is likewise nonlinear in its arguments and thus corresponds to an RKHS much richer than $\mathbb{R}^Q$. Specifically, we use a Gaussian kernel

$$k(\mathbf{p}, \mathbf{p}') \leftarrow \exp\left(-\frac{1}{2} \big\| \mathbf{p} - \mathbf{p}' \big\|_{\boldsymbol{\Lambda}^{-2}}^2 \right), \tag{12}$$

where positive definite matrix bandwidth $\boldsymbol{\Lambda} \in \mathbb{R}^{Q \times Q}$ controls the length scales in $\mathbf{p}$ over which the estimator $\widehat{\mathbf{x}}$ smooths and $\|\cdot\|_{\boldsymbol{\Gamma}} \equiv \big\| \boldsymbol{\Gamma}^{1/2}(\cdot) \big\|_2$ is a weighted ℓ^2-norm with PSD matrix weights $\boldsymbol{\Gamma}$. We use a Gaussian kernel over other candidates because it is a *universal kernel*, meaning weighted sums of the form $\sum_{n=1}^{N} a_n k(\cdot, \mathbf{p}_n)$ can approximate $\mathcal{L}^2$ functions to arbitrary accuracy for N sufficiently large [44].

Interestingly, the RKHS associated with Gaussian kernel (12) is infinite-dimensional. Thus, Gaussian kernel regression can be interpreted as first "lifting" via a nonlinear *feature map* $\mathbf{z} : \mathbb{R}^Q \mapsto \bar{\mathbb{H}}$ each $\mathbf{p}$ into an infinite-dimensional *feature* $\mathbf{z}(\mathbf{p}) = k(\cdot, \mathbf{p}) \in \bar{\mathbb{H}}$, and then performing regularized affine regression on the features via dot products of the form $\langle k(\cdot, \mathbf{p}), k(\cdot, \mathbf{p}') \rangle_{\bar{\mathbb{H}}} = k(\mathbf{p}', \mathbf{p})$. From this perspective, the challenges of nonlinear estimation via likelihood models are avoided because we *select* (through the choice of kernel) characteristics of the nonlinear dependence that we wish to model and need only *estimate* via (8) the linear dependence of each entry in $\widehat{\mathbf{x}}$ on the corresponding features.

IV. Bias and Covariance Analysis

This section presents expressions for the bias and covariance of Gaussian PERK estimator $\widehat{\mathbf{x}}(\cdot)$, conditioned on object parameters $\mathbf{x}, \boldsymbol{\nu}$. We focus on these conditional statistics to enable study of estimator performance as $\mathbf{x}, \boldsymbol{\nu}$ are varied. Though not mentioned explicitly hereafter, both expressions treat the

[3]We present our methodology assuming pure-real patterns $\mathbf{p}$ and estimators $\widehat{\mathbf{x}}$ for simplicity and to maintain consistency with experiments, in which we choose to use magnitude images for unrelated reasons (see §VI.A for details). It is straightforward to generalize Theorem 1 for complex-valued kernels and thereby address the cases of complex patterns and/or estimators.

training sample $\{(\mathbf{x}_1, \mathbf{p}_1), \ldots, (\mathbf{x}_N, \mathbf{p}_N)\}$ and regularization parameters $\rho_1, \ldots, \rho_L$ as fixed.

A. Conditional Bias

The conditional bias of $\widehat{\mathbf{x}} \equiv \widehat{\mathbf{x}}(\boldsymbol{\alpha}, \boldsymbol{\nu})$ is written as

$$\begin{aligned}\operatorname{bias}(\widehat{\mathbf{x}}|\mathbf{x}, \boldsymbol{\nu}) &:= \mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\widehat{\mathbf{x}}(\boldsymbol{\alpha}, \boldsymbol{\nu})) - \mathbf{x} \\ &= \mathbf{R}\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\mathbf{k}(\boldsymbol{\alpha}, \boldsymbol{\nu})) + (\mathbf{m}_{\mathbf{x}} - \mathbf{x}), \end{aligned} \tag{13}$$

where $\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\cdot)$ denotes expectation with respect to $\boldsymbol{\alpha} := |\mathbf{y}|$ and conditioned on $\mathbf{x}$, $\boldsymbol{\nu}$. Here, the lth row of $\mathbf{R} \in \mathbb{R}^{L\times N}$ and lth entry of regressand sample mean $\mathbf{m}_{\mathbf{x}} \in \mathbb{R}^L$ respectively are $\boldsymbol{x}_l^{\mathsf{T}}\mathbf{M}(\mathbf{M}\mathbf{K}\mathbf{M} + N\rho_l\mathbf{I}_N)^{-1}$ and $\frac{1}{N}\boldsymbol{x}_l^{\mathsf{T}}\mathbf{1}_N$ for $l \in \{1, \ldots, L\}$. To proceed analytically, we make two mild assumptions. First, we assume that $\mathbf{y} \sim \mathbb{C}\mathcal{N}(\mathbf{0}_D, \boldsymbol{\Sigma})$ has sufficiently high signal-to-noise ratio (SNR) such that its complex modulus $\boldsymbol{\alpha}$ is approximately Gaussian-distributed. We specifically consider the typical case where covariance matrix $\boldsymbol{\Sigma}$ is diagonal with diagonal entries $\sigma_1^2, \ldots, \sigma_D^2$, in which case measurement amplitude conditional distribution $\mathsf{p}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}$ is simply approximated as $\mathsf{p}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}} \leftarrow \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^D$ has dth coordinate $\sqrt{|s_d(\mathbf{x}, \boldsymbol{\nu})|^2 + \sigma_d^2}$ for $d \in \{1, \ldots, D\}$ [45]. Second, we assume that the Gaussian kernel bandwidth matrix $\boldsymbol{\Lambda}$ has the block diagonal structure

$$\boldsymbol{\Lambda} \leftarrow \begin{bmatrix} \boldsymbol{\Lambda}_{\boldsymbol{\alpha}} & \mathbf{0}_{D\times K} \\ \mathbf{0}_{K\times D} & \boldsymbol{\Lambda}_{\boldsymbol{\nu}}. \end{bmatrix} \tag{14}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\alpha}} \in \mathbb{R}^{D\times D}$ and $\boldsymbol{\Lambda}_{\boldsymbol{\nu}} \in \mathbb{R}^{K\times K}$ are positive definite. With these simplifying assumptions, the nth entry of the expectation in (13) is well approximated as $\left[\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\mathbf{k}(\boldsymbol{\alpha}, \boldsymbol{\nu}))\right]_n$

$$\begin{aligned} &= \int_{\mathbb{R}^D} e^{-\frac{1}{2}\|\mathbf{p}-\mathbf{p}_n\|^2_{\boldsymbol{\Lambda}^{-2}}} \mathsf{p}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\boldsymbol{\alpha}|\mathbf{x}, \boldsymbol{\nu})\, \mathrm{d}\,\boldsymbol{\alpha} \\ &\approx \frac{e^{-\frac{1}{2}\|\boldsymbol{\nu}-\boldsymbol{\nu}_n\|^2_{\boldsymbol{\Lambda}_{\boldsymbol{\nu}}^{-2}}}}{\sqrt{(2\pi)^D \det(\boldsymbol{\Sigma})}} \int_{\mathbb{R}^D} e^{-\frac{1}{2}\left(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_n\|^2_{\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}} + \|\boldsymbol{\alpha}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\right)} \mathrm{d}\,\boldsymbol{\alpha} \\ &= \frac{e^{-\frac{1}{2}\left(\|\boldsymbol{\nu}-\boldsymbol{\nu}_n\|^2_{\boldsymbol{\Lambda}_{\boldsymbol{\nu}}^{-2}} + \|\boldsymbol{\mu}-\boldsymbol{\alpha}_n\|^2_{\left(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}\boldsymbol{\Sigma}+\mathbf{I}_D\right)^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}}\right)}}{\sqrt{\det\left(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}\boldsymbol{\Sigma} + \mathbf{I}_D\right)}}, \end{aligned} \tag{15}$$

where $\det(\cdot)$ denotes determinant and the Gaussian integral follows after completing the square of the integrand's exponent. It is clear from (15) that as $\boldsymbol{\Sigma} \to \mathbf{0}_{D\times D}$ for fixed $\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}$, $\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\mathbf{k}(\boldsymbol{\alpha}, \boldsymbol{\nu})) \to \mathbf{k}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and therefore

$$\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\widehat{\mathbf{x}}(\boldsymbol{\alpha}, \boldsymbol{\nu})) \to \widehat{\mathbf{x}}\left(\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\boldsymbol{\alpha}), \boldsymbol{\nu}\right) \equiv \widehat{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \tag{16}$$

which perhaps surprisingly means that the conditional bias asymptotically approaches the noiseless conditional estimation error $\widehat{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \mathbf{x}$ despite $\widehat{\mathbf{x}}$ being nonlinear in $\boldsymbol{\alpha}$.

B. Conditional Covariance

The conditional covariance of $\widehat{\mathbf{x}} \equiv \widehat{\mathbf{x}}(\boldsymbol{\alpha}, \boldsymbol{\nu})$ is written as

$$\begin{aligned}\operatorname{cov}(\widehat{\mathbf{x}}|\mathbf{x}, \boldsymbol{\nu}) &:= \mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}\left(\left(\widehat{\mathbf{x}} - \mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\widehat{\mathbf{x}})\right)\left(\widehat{\mathbf{x}} - \mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\widehat{\mathbf{x}})\right)^{\mathsf{T}}\right) \\ &= \mathbf{R}\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}\left(\widetilde{\mathbf{k}}(\boldsymbol{\alpha}, \boldsymbol{\nu})\widetilde{\mathbf{k}}(\boldsymbol{\alpha}, \boldsymbol{\nu})^{\mathsf{T}}\right)\mathbf{R}^{\mathsf{T}}, \end{aligned} \tag{17}$$

where $\widetilde{\mathbf{k}}(\boldsymbol{\alpha}, \boldsymbol{\nu}) := \mathbf{k}(\boldsymbol{\alpha}, \boldsymbol{\nu}) - \mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}(\mathbf{k}(\boldsymbol{\alpha}, \boldsymbol{\nu}))$. To proceed analytically, we take the same high-SNR and block-diagonal bandwidth assumptions as in §IV.A. Then after straightforward manipulations similar to those yielding (15), the (n, n')th entry of the expectation in (17) is well approximated as

$$\begin{aligned} &\left[\mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}\left(\widetilde{\mathbf{k}}(\boldsymbol{\alpha}, \boldsymbol{\nu})\widetilde{\mathbf{k}}(\boldsymbol{\alpha}, \boldsymbol{\nu})^{\mathsf{T}}\right)\right]_{n,n'} \\ &\quad = e^{-\frac{1}{2}\left(\|\boldsymbol{\nu}-\boldsymbol{\nu}_n\|^2_{\boldsymbol{\Lambda}_{\boldsymbol{\nu}}^{-2}} + \|\boldsymbol{\nu}-\boldsymbol{\nu}_{n'}\|^2_{\boldsymbol{\Lambda}_{\boldsymbol{\nu}}^{-2}}\right)} \\ &\qquad \times \left(\frac{e^{-\frac{1}{2}\left(\|\widetilde{\boldsymbol{\alpha}}_n - \widetilde{\boldsymbol{\alpha}}_{n'}\|^2_{\boldsymbol{\Delta}(0)} + \|\widetilde{\boldsymbol{\alpha}}_n + \widetilde{\boldsymbol{\alpha}}_{n'}\|^2_{\boldsymbol{\Delta}(2)}\right)}}{\sqrt{\det\left(2\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}\boldsymbol{\Sigma} + \mathbf{I}_D\right)}} \right. \\ &\qquad \left. - \frac{e^{-\frac{1}{2}\left(\|\widetilde{\boldsymbol{\alpha}}_n - \widetilde{\boldsymbol{\alpha}}_{n'}\|^2_{\boldsymbol{\Delta}(1)} + \|\widetilde{\boldsymbol{\alpha}}_n + \widetilde{\boldsymbol{\alpha}}_{n'}\|^2_{\boldsymbol{\Delta}(1)}\right)}}{\det\left(\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}\boldsymbol{\Sigma} + \mathbf{I}_D\right)} \right), \end{aligned} \tag{18}$$

where $\widetilde{\boldsymbol{\alpha}}_n := \boldsymbol{\mu} - \boldsymbol{\alpha}_n$ and $\boldsymbol{\Delta}(t) := \frac{1}{2}\left(t\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}\boldsymbol{\Sigma} + \mathbf{I}_D\right)^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\alpha}}^{-2}$ for $t \in \mathbb{N}$. The emergence of $\widetilde{\boldsymbol{\alpha}}_n \pm \widetilde{\boldsymbol{\alpha}}_{n'}$ terms in (18) show that the conditional covariance (unlike the conditional bias) is directly influenced not only by the individual expected test point distances to each of the training points $\widetilde{\boldsymbol{\alpha}}_1, \ldots, \widetilde{\boldsymbol{\alpha}}_N$ but also by the local training point sampling density.

V. Implementation Considerations

This section focuses on important practical implementation issues. §V.A discusses a conceptually intuitive approximation of PERK estimator (11) that in many problems can significantly improve computational performance. §V.B describes strategies for data-driven model selection.

A. A Kernel Approximation

In practical problems with even moderately large ambient dimension P, the necessarily large number of training samples N complicates storage of (dense) $N \times N$ Gram matrix $\mathbf{K}$. Using a kernel approximation can mitigate storage and processing issues. Here we employ *random Fourier features* [46], a recent method for approximating translation-invariant kernels having form $k(\mathbf{p}, \mathbf{p}') \equiv k(\mathbf{p} - \mathbf{p}')$. This subsection reviews the main result of [46] for the purpose of constructing an intuitive and computationally efficient approximation of (11).

The strategy of [46] is to construct independent probability distributions $\mathsf{p}_{\mathbf{v}}$ and p_s associated with random $\mathbf{v} \in \mathbb{R}^P$ and random $s \in \mathbb{R}$ as well as a function (that is parameterized by $\mathbf{p}$) $\tilde{z}(\cdot, \cdot; \mathbf{p}) : \mathbb{R}^P \times \mathbb{R} \times \mathbb{R}^P \mapsto \mathbb{R}$, such that

$$\mathsf{E}_{\mathbf{v},s}\left(\tilde{z}(\mathbf{v}, s; \mathbf{p})\tilde{z}(\mathbf{v}, s; \mathbf{p}')\right) = k(\mathbf{p} - \mathbf{p}'), \tag{19}$$

where $\mathsf{E}_{\mathbf{v},s}(\cdot)$ denotes expectation with respect to $\mathsf{p}_{\mathbf{v}}\mathsf{p}_s$. When such a construction exists, one can build approximate feature maps $\tilde{\mathbf{z}}$ by concatenating and normalizing evaluations of $\tilde{z}$ on Z samples $\{(\mathbf{v}_1, s_1), \ldots, (\mathbf{v}_Z, s_Z)\}$ of $(\mathbf{v}, s)$ (drawn jointly albeit independently), to produce approximate features

$$\tilde{\mathbf{z}}(\mathbf{p}) := \sqrt{\frac{2}{Z}}\left[\tilde{z}(\mathbf{v}_1, s_1; \mathbf{p}), \ldots, \tilde{z}(\mathbf{v}_Z, s_Z; \mathbf{p})\right]^{\mathsf{T}} \tag{20}$$

for any $\mathbf{p}$. Then by the strong law of large numbers,

$$\lim_{Z\to\infty} \langle \tilde{\mathbf{z}}(\mathbf{p}), \tilde{\mathbf{z}}(\mathbf{p}')\rangle_{\mathbb{R}^Z} \overset{a.s.}{\to} k(\mathbf{p}, \mathbf{p}') \qquad \forall \mathbf{p}, \mathbf{p}' \tag{21}$$

which, in conjunction with strong performance guarantees for finite Z [46], [47], justifies interpreting $\tilde{\mathbf{z}}$ as an approximate (and now finite-dimensional) feature map.

We use the Fourier construction of [46] that assigns $\tilde{z}(\mathbf{v}, s; \mathbf{p}) \leftarrow \cos\left(2\pi\left(\mathbf{v}^{\mathsf{T}}\mathbf{p} + s\right)\right)$. If also $\mathsf{p}_s \leftarrow \text{unif}(0, 1)$, then $\mathsf{E}_{\mathbf{v},s}\left(\tilde{z}(\mathbf{v}, s; \mathbf{p})\tilde{z}\left(\mathbf{v}, s; \mathbf{p}'\right)\right)$ simplifies to

$$\int_{\mathbb{R}^P} \cos\left(2\pi\,\mathbf{v}^{\mathsf{T}}\left(\mathbf{p} - \mathbf{p}'\right)\right)\mathsf{p}_{\mathbf{v}}(\mathbf{v})\;\mathrm{d}\,\mathbf{v}. \tag{22}$$

For symmetric $\mathsf{p}_{\mathbf{v}}$, (22) exists [48] and is a Fourier transform. Thus choosing $\mathsf{p}_{\mathbf{v}} \leftarrow \mathcal{N}\left(\mathbf{0}_P, (2\pi\,\mathbf{\Lambda})^{-2}\right)$ satisfies (19) for Gaussian kernel (12), where $\mathbf{0}_P \in \mathbb{R}^P$ is a vector of zeros.

Sampling $\mathsf{p}_{\mathbf{v}}, \mathsf{p}_s$ Z times and subsequently constructing $\tilde{\mathbf{Z}} := \left[\tilde{\mathbf{z}}(\mathbf{p}_1), \ldots, \tilde{\mathbf{z}}(\mathbf{p}_N)\right] \in \mathbb{R}^{Z\times N}$ via repeated evaluations of (20) gives for $Z \ll N$ a low-rank approximation $\tilde{\mathbf{Z}}^{\mathsf{T}}\tilde{\mathbf{Z}}$ of Gram matrix $\mathbf{K}$. Substituting this approximation into (11) and applying the matrix inversion lemma [49] yields

$$\widehat{x}_l(\cdot) \leftarrow m_{x_l} + \mathbf{c}_{\mathbf{z}x_l}^{\mathsf{T}}(\mathbf{C}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} + \rho_l\mathbf{I}_Z)^{-1}(\tilde{\mathbf{z}}(\cdot) - \mathbf{m}_{\tilde{\mathbf{z}}}), \tag{23}$$

where $m_{x_l} := \frac{1}{N}\boldsymbol{x}_l^{\mathsf{T}}\mathbf{1}_N$ and $\mathbf{m}_{\tilde{\mathbf{z}}} := \frac{1}{N}\tilde{\mathbf{Z}}\mathbf{1}_N$ are sample means; and $\mathbf{c}_{\mathbf{z}x_l} := \frac{1}{N}\tilde{\mathbf{Z}}\mathbf{M}\boldsymbol{x}_l$ and $\mathbf{C}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} := \frac{1}{N}\tilde{\mathbf{Z}}\mathbf{M}\tilde{\mathbf{Z}}^{\mathsf{T}}$ are sample covariances. Estimator (23) is an affine minimum mean-squared error estimator on the approximate features, and illustrates that Gaussian PERK via estimator (11) is asymptotically (in Z) equivalent to regularized affine regression after nonlinear, high-dimensional feature mapping.

B. Tuning Parameter Selection

This subsection proposes guidelines for data-driven selection of user-selectable parameters. Our goal here is to use problem intuition to automatically choose as many tuning parameters as possible, thereby leaving as few parameters as possible to manual selection. In this spirit, we focus on "online" model selection, where one chooses tuning parameters for training the estimator $\widehat{\mathbf{x}}(\cdot)$ *after* acquiring (unlabeled) real test data. This online approach can be considered a form of *transductive learning* [50, Ch. 8] since we train our estimator with knowledge of unlabeled test data in addition to labeled training data. Observe that since many voxel-wise separable MRI parameter estimation problems are comparatively low-dimensional, PERK estimators can often be quickly trained using only a moderate number of simulated training examples; in fact, training can in some problems take comparable or even less time than evaluating the PERK estimator on full-volume high-resolution measurement images. For these reasons, online PERK model selection is often practical.

1) Choosing Sampling Distribution: For reasonable PERK performance, it is important to choose the joint distribution of latent and known parameters $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}$ such that latent parameters can be estimated precisely over the joint distribution's support $\mathsf{supp}\left(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}\right)$. For continuously differentiable magnitude signal model $\boldsymbol{\mu}$, we quantify precision at a single point $(\mathbf{x}, \boldsymbol{\nu})$ using the Fisher information matrix

$$\begin{aligned}\mathbf{F}(\mathbf{x}, \boldsymbol{\nu}) &:= \mathsf{E}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}\left(\left(\nabla_{\mathbf{x}}\log\mathsf{p}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}\right)^{\mathsf{T}}\nabla_{\mathbf{x}}\log\mathsf{p}_{\boldsymbol{\alpha}|\mathbf{x},\boldsymbol{\nu}}\right)\\ &\approx \left(\nabla_{\mathbf{x}}\boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\nu})\right)^{\mathsf{T}}\mathbf{\Sigma}^{-1}\nabla_{\mathbf{x}}\boldsymbol{\mu}(\mathbf{x}, \boldsymbol{\nu})\end{aligned} \tag{24}$$

where $\nabla_{\mathbf{x}}(\cdot)$ denotes row gradient with respect to $\mathbf{x}$ and the approximation holds well for moderately high-SNR measurements [45]. When it exists, the inverse of $\mathbf{F}(\mathbf{x}, \boldsymbol{\nu})$ provides a lower-bound on the conditional covariance of any unbiased estimator of $\mathbf{x}$ [51]. For good performance, it is thus reasonable to ensure $\mathbf{F}(\mathbf{x}, \boldsymbol{\nu})$ is well-conditioned over $\mathsf{supp}\left(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}\right)$.

There are many strategies one could employ to control the condition number of $\mathbf{F}(\mathbf{x}, \boldsymbol{\nu})$ over $\mathsf{supp}\left(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}\right)$. In our experiments, we used data [19] from acquisitions designed to *minimize* a cost function related to the *maximum* of $\mathbf{F}^{-1}(\mathbf{x}, \boldsymbol{\nu})$ over bounded latent and known parameter ranges of interest (§VI.A provides application-specific details). We then assigned $\mathsf{supp}\left(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}\right)$ to coincide with the support of these acquisition design parameter ranges of interest. Assessing worst-case imprecision via the conservative minimax criterion is appropriate here because point-wise poor conditioning at any $(\mathbf{x}, \boldsymbol{\nu}) \in \mathsf{supp}\left(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}\right)$ can induce PERK estimation error over larger subsets of $\mathsf{supp}\left(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}\right)$.

If many separate prior parameter estimates are available, one can estimate the particular shape of $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}$ empirically and then clip and renormalize $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}$ so as to assign nonzero probability only within an appropriate support. When prior estimates are unavailable, it may in certain problems be reasonable to instead assume a separable distributional structure $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}} \equiv \mathsf{p}_{\mathbf{x}}\mathsf{p}_{\boldsymbol{\nu}}$ in which case one can still estimate $\mathsf{p}_{\boldsymbol{\nu}}$ empirically but must set $\mathsf{p}_{\mathbf{x}}$ manually based on typical ranges of latent parameters.

2) Choosing Regularization Parameters: As presented, PERK estimator (11) and its approximation (23) leave freedom to select different regularization parameters $\rho_1, \ldots, \rho_L$ for estimating each of the L latent parameters. However, the respective unitless matrices $\mathbf{MKM}$ and $\mathbf{C}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$ whose condition numbers are influenced by $\rho_1, \ldots, \rho_L$ do not vary with l. Thus it is reasonable to assign each $\rho_l \leftarrow \rho\ \forall l \in \{1, \ldots, L\}$ some fixed $\rho > 0$. This simplification significantly reduces training computation to just one rather than L large matrix inversions. We select the scalar regularization parameter ρ using the holdout process described in §S.II.

3) Choosing Kernel Bandwidth: It is desirable to choose the Gaussian kernel's bandwidth matrix $\mathbf{\Lambda}$ such that PERK estimates are invariant to the overall scale of test data. We use (after observing test data, and for both training and testing)

$$\mathbf{\Lambda} \leftarrow \lambda\,\mathrm{diag}\left(\left[\mathbf{m}_{\boldsymbol{\alpha}}^{\mathsf{T}}, \mathbf{m}_{\boldsymbol{\nu}}^{\mathsf{T}}\right]^{\mathsf{T}}\right), \tag{25}$$

where $\mathbf{m}_{\boldsymbol{\alpha}} \in \mathbb{R}^D$ and $\mathbf{m}_{\boldsymbol{\nu}} \in \mathbb{R}^K$ are sample means across voxels of magnitude test image data and known parameters, respectively; and $\mathrm{diag}(\cdot)$ assigns its argument to the diagonal entries of an otherwise zero matrix. We select the only scalar bandwidth parameter $\lambda > 0$ using holdout as well.

VI. Experimentation

This section demonstrates PERK for quantifying MR relaxation parameters T_1 and T_2, a well-studied application. We studied this relatively simple problem instead of the more complicated problems that motivated our method because we had access to reference T_1, T_2 phantom NMR measurements [52] for external validation and because it is easier to validate PERK estimates against gold-standard

grid search estimates in problems involving few unknowns. §VI.A describes implementation details that were fixed in all simulations and experiments. §VI.B studies estimator statistics in numerical simulations. §VI.C and §VI.D respectively compare PERK performance in phantom and *in vivo* experiments.

A. Methods

In all simulations and experiments, we used data arising from a fast acquisition [19] consisting of two spoiled gradient-recalled echo (SPGR) [53] and one dual-echo steady-state (DESS) [54] scans. Since each SPGR (DESS) scan generates one (two) signal(s) per excitation, this acquisition yielded $D \leftarrow 4$ datasets. We fixed scan parameters to be identical to those in [19], wherein repetition times and flip angles were optimized for precise T_1 and T_2 estimation in cerebral tissue at 3T field strength [19] and echo times were fixed across scans. We used standard magnitude[4] SPGR and DESS signal models expressed as a function of four free parameters per voxel: flip angle spatial variation (due to transmit field inhomogeneity) κ; longitudinal and transverse relaxation time constants T_1 and T_2; and a pure-real proportionality constant M_0. We assumed prior knowledge of $K \leftarrow 1$ known parameter $\boldsymbol{\nu} \leftarrow \kappa$ (in experiments, through separate acquisition and estimation of flip angle scaling maps) and collected the remaining $L \leftarrow 3$ latent parameters as $\mathbf{x} \leftarrow [M_0, T_1, T_2]^{\mathsf{T}}$.

We used the same PERK training and testing process across all simulations and experiments. We assumed a separable prior distribution $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}} \leftarrow \mathsf{p}_{M_0,T_1,T_2,\kappa} \equiv \mathsf{p}_{M_0}\mathsf{p}_{T_1}\mathsf{p}_{T_2}\mathsf{p}_{\kappa}$ and estimated flip angle scaling marginal distribution p_{κ} from known κ map voxels via kernel density estimation (implemented using the built-in MATLAB® function `fitdist` with default options). To match the scaling of training and test data, we set M_0 marginal distribution $\mathsf{p}_{M_0} \leftarrow \text{unif}\left(2.2 \times 10^{-16}, u\right)$, with u set as $6.67\times$ the maximum value of magnitude test data. We chose the supports of T_1, T_2 marginal distributions $\mathsf{p}_{T_1} \leftarrow \text{logunif}(400, 2000)$ms, $\mathsf{p}_{T_2} \leftarrow \text{logunif}(40, 200)$ms and clipped the support of p_{κ} to assign nonzero probability only within $[0.5, 2]$ such that these supports coincided with the supports over which [19] optimized the acquisition. We assumed noise covariance $\boldsymbol{\Sigma}$ of form $\sigma^2\mathbf{I}_4$ (as in [19]) and estimated the (spatially invariant) noise variance σ^2 from Rayleigh-distributed regions of magnitude test data, using estimators described in [55]. We sampled $N \leftarrow 10^5$ latent and known parameter realizations from these distributions and evaluated SPGR and DESS signal models to generate corresponding noiseless measurements. After adding complex Gaussian noise realizations, we concatenated the (Rician) magnitude of these noisy measurements with known parameter realizations to construct pure-real regressors. We separately selected and then held fixed free parameters $\lambda \leftarrow 2^{0.6}$ and $\rho \leftarrow 2^{-41}$ via a simple holdout process in simulation, described in §S.II. We set Gaussian kernel bandwidth matrix $\boldsymbol{\Lambda}$ from test data via (25). We sampled $\mathbf{v}, s$ $Z \leftarrow 10^3$ times to construct approximate feature map $\tilde{\mathbf{z}}$. For each latent parameter $l \leftarrow \{1, \ldots, L\}$, we applied $\tilde{\mathbf{z}}$ to training data; computed sample means $m_{x_l}, \mathbf{m}_{\tilde{\mathbf{z}}}$ and sample covariances $\mathbf{c}_{\mathbf{z}x_l}, \mathbf{C}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$; and evaluated (23) on test image data and the known flip angle scaling map on a per-voxel basis.

We evaluated PERK latent parameter estimates against maximum-likelihood (ML) estimates computed via two well-suited algorithms that we describe here in turn. We first implemented a grid search estimator accelerated by the variable projection method (VPM) [56], a popular technique that has been used in many QMRI algorithms and applications (see [7]–[9], [11]–[13], [16], [18], [19], [57], [58]). Following [19], we clustered flip angle scaling map voxels into 20 clusters via k-means++ [59] and used each of the 20 cluster means along with 500 T_1 and T_2 values logarithmically spaced between $\left(10^{1.5}, 10^{3.5}\right)$ and $\left(10^{0.5}, 10^{3}\right)$ to compute 20 dictionaries, each consisting of $250,000$ signal vectors (fewer clusters introduced noticeable errors in experiments). Iterating over clusters, we generated each cluster's dictionary and applied VPM and grid search over magnitude image data voxels assigned to that cluster.

We also compared PERK to iterative ML optimization via a preconditioned variant of the classical gradient projection method (PGPM) [60]. We designed the preconditioner as the inverse of a positive definite diagonal majorizer of the negative log-likelihood cost function's Hessian matrix, updated for the first five iterations and fixed thereafter. We employed a diagonal preconditioner to retain the linear convergence rate guarantees of GPM [61] yet accelerate practical performance. We initialized PGPM via conventional method-of-moments estimators of M_0, T_1 from 2 SPGR scans [62] and T_2 from 1 DESS scan [54] (the method-of-moments T_2 estimator is strongly biased). We used the MATLAB® Symbolic Toolbox to generate cumbersome but analytical expressions for the gradient and Hessian of the magnitude SPGR and DESS signal models. At each PGPM iteration, we used these expressions to compute a preconditioned descent direction, update the iterate, and project each voxel's T_1 and T_2 iterate to lie within $[100, 3000]$ms and $[10, 700]$ms, respectively. We continued iterations until the convergence criterion

$$\left\|\boldsymbol{\Omega}^{-1}\left(\mathbf{X}^{(i)} - \mathbf{X}^{(i-1)}\right)\right\|_{\mathrm{F}} < 10^{-7}\left\|\boldsymbol{\Omega}^{-1}\left(\mathbf{X}^{(i-1)}\right)\right\|_{\mathrm{F}} \quad (26)$$

was satisfied, where $\mathbf{X}$ collects latent parameter voxels in its columns, $(\cdot)^{(i)}$ denotes the ith iterate, $\boldsymbol{\Omega} := \text{diag}\left(\text{med}\left(\mathbf{X}^{(0)}\right)\right)$ is a fixed latent parameter weighting matrix, and $\text{med}(\cdot)$ takes the median across the columns of its argument.

To ensure monotone local convergence in cost, we implemented PGPM to include a simple step-halving line search at each iteration. In early experiments however, we observed even in simulation and even with preconditioning that attempting to update all voxels simultaneously using a single line search resulted in large errors due to excessive step-halving and subsequent early termination of iterations. To circumvent separate line searches for every voxel, we first clustered latent parameter initializations and flip angle scaling map voxels into

[4]Standard complex DESS signal models depend on a fifth free parameter associated with phase accrual due to off-resonance effects. Because the first and second DESS signals depend differently on off-resonance phase accrual [19], off-resonance related phase (unlike signal loss) cannot be collected into the (now complex) proportionality constant. To avoid (separate or joint) estimation of an off-resonance field map, we followed [19] and used magnitude SPGR and DESS signal models. We accounted for consequently Rician-distributed noise in magnitude image data during training.

TABLE I

SAMPLE MEANS ± SAMPLE STANDARD DEVIATIONS (RMSES) OF VPM, PGPM, AND PERK T_1, T_2 ESTIMATES, COMPUTED IN SIMULATION OVER 7810 WM-LIKE AND 9162 GM-LIKE VOXELS. EACH SAMPLE STATISTIC IS ROUNDED OFF TO THE HIGHEST PLACE VALUE OF ITS (UNREPORTED) STANDARD ERROR, COMPUTED VIA FORMULAS IN [63]. ALL VALUES ARE REPORTED IN MILLISECONDS

	Truth	VPM	PGPM	PERK
WM T_1	832	832.1 ± 17.2 (17.2)	832.1 ± 16.2 (16.2)	833.0 ± 16.5 (16.5)
GM T_1	1331	1331.5 ± 31.1 (31.1)	1331.2 ± 29.7 (29.7)	1332.1 ± 30.4 (30.4)
WM T_2	79.6	79.61 ± 0.988 (0.988)	79.60 ± 0.952 (0.952)	79.46 ± 0.978 (0.989)
GM T_2	110.	110.02 ± 1.40 (1.40)	110.02 ± 1.35 (1.35)	109.91 ± 1.35 (1.35)

50 clusters and then ran PGPM separately on each cluster (fewer clusters reintroduced early stopping).

We performed all simulations and experiments running MATLAB® R2013a on a 3.5GHz desktop computer equipped with 32GB RAM. Because our experiments use a single slice of image data, we report PERK training and testing times separately and note that only the latter time would scale linearly with the number of voxels (the former would scale negligibly due only to online model selection). In the interest of reproducible research, code and data will be freely available at https://gitlab.eecs.umich.edu/fessler/qmri.

B. Numerical Simulations

We assigned typical T_1, T_2 values in white matter (WM) and grey matter (GM) at 3T [64] to the discrete anatomy of the 81st slice of the BrainWeb digital phantom [65] to produce ground truth M_0, T_1, T_2 maps. We simulated 217×181 noiseless single-coil SPGR and DESS image data, modeling (and then assuming as known) 20% flip angle spatial variation κ. We corrupted noiseless datasets with additive complex Gaussian noise to yield noisy complex datasets with SNR ranging from 94-154 in WM and 82-154 in GM, where SNR is defined

$$\mathsf{SNR}(\tilde{\mathbf{y}}, \tilde{\boldsymbol{\epsilon}}) := \|\tilde{\mathbf{y}}\|_2 / \|\tilde{\boldsymbol{\epsilon}}\|_2 \tag{27}$$

for image data voxels $\tilde{\mathbf{y}}$ and noise voxels $\tilde{\boldsymbol{\epsilon}}$ corresponding to a region of interest (ROI) within a single SPGR/DESS dataset. We estimated M_0, T_1, T_2 from noisy magnitude images and known κ maps using VPM, PGPM, and PERK. VPM took 791s; PGPM took 1821s; and PERK training and testing respectively took 3.6s and 1.5s.

Table I compares sample statistics of VPM, PGPM, and PERK T_1, T_2 estimates, computed over 7810 WM-like and 9162 GM-like voxels (§S.IV presents corresponding images and M_0 sample statistics). Overall, all three methods achieve excellent performance. PERK estimates are slightly more precise but slightly less accurate than gold-standard VPM estimates. Results suggest that at least in WM- and GM-like voxels, PGPM is capable of descending the ML cost towards a desirable solution; in fact, PGPM achieves slightly better precision than either VPM or PERK. All three methods exhibit comparable root mean squared errors (RMSEs).

C. Phantom Experiments

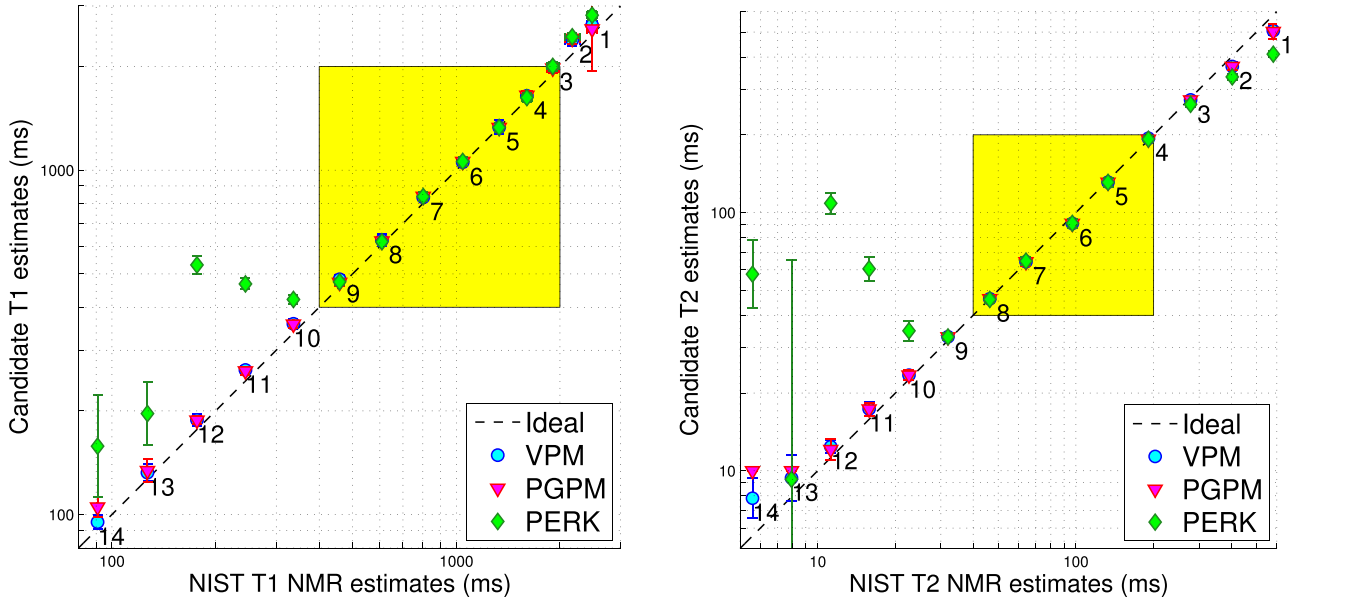
Phantom experiments used datasets from fast coronal scans of a High Precision Devices® MR system phantom T_2 array acquired on a 3T GE Discovery™ scanner with an 8-channel receive head array. This acquisition consisted of: two SPGR scans with 5, 15° flip angles and 12.2, 12.2ms repetition times; one DESS scan with 30° flip angle and 17.5ms repetition time; and two Bloch-Siegert (BS) scans [26] (for separate flip angle scaling κ estimation). Nominal flip angles were achieved by scaling a 2cm slab-selective Shinnar-Le Roux RF excitation [66] of duration 1.28ms and time-bandwidth product 4. All scans collected fully-sampled 3D Cartesian data using 4.67ms echo times with a $256 \times 256 \times 8$ matrix over a $24 \times 24 \times 4\text{cm}^3$ field of view. Scan time totaled 3m17s. The scan room temperature was recorded as 293K once at the beginning of the exam. Further acquisition details are reported in [19].

For each SPGR, DESS, and BS dataset, we reconstructed raw coil images via 3D Fourier transform and subsequently processed only one image slice centered within the excitation slab. We combined SPGR and DESS coil images using a natural extension of [67] to the case of multiple datasets. We similarly (but separately) combined BS coil images and estimated κ maps by normalizing and calibrating regularized transmit field estimates [68] from complex coil-combined BS images. We estimated M_0, T_1, T_2 from magnitude SPGR/DESS images and κ maps using VPM, PGPM, and PERK. VPM took 928s; PGPM took 1257s; and PERK training and testing respectively took 4.2s and 1.9s.

Fig. 1 compares sample means and sample standard deviations computed within ROIs of VPM, PGPM, and PERK T_1, T_2 estimates against nuclear magnetic resonance (NMR) reference measurements reported at 293.00K from the National Institute for Standards of Technology (NIST) [52]. Yellow box boundaries indicate projections of the PERK sampling distribution's support $\mathsf{supp}(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}})$. ROI labels correspond with vial markers depicted in images presented in §S.V.A. Within $\mathsf{supp}(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}})$, corresponding tables demonstrate that VPM, PGPM, and PERK estimates agree excellently with each other and reasonably with NMR measurements. We do not expect good PERK performance outside $\mathsf{supp}(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}})$ and indeed observe poor ability to extrapolate. As discussed in §V.B.1 and demonstrated in §S.V.B, expanding $\mathsf{supp}(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}})$ well beyond the acquisition design parameter range of interest can substantially reduce PERK performance for typical T_1, T_2 WM and GM values.

D. In vivo Experiments

In vivo experiments used datasets from axial scans of a healthy volunteer acquired with a 32-channel Nova Medical®



T_1	NMR	VPM	PGPM	PERK
V4	1604 ± 7.2	1645 ± 48	1649 ± 48	1626 ± 46
V5	1332 ± 0.8	1335 ± 61	1331 ± 41	$1332 \pm 40.$
V6	1044 ± 3.2	1055 ± 28	$1060. \pm 29$	1061 ± 29
V7	801.7 ± 1.70	834 ± 21	$840. \pm 23$	839 ± 23
V8	608.6 ± 1.03	627 ± 25	623 ± 12	$620. \pm 13$

T_2	NMR	VPM	PGPM	PERK
V4	190.94 ± 0.011	194 ± 5.5	192.4 ± 5.2	192.5 ± 4.9
V5	133.27 ± 0.073	131.2 ± 5.3	131 ± 5.5	131 ± 5.5
V6	96.89 ± 0.049	90.8 ± 3.5	90.8 ± 3.5	90.9 ± 3.5
V7	64.07 ± 0.034	64.6 ± 2.2	64.5 ± 2.1	65.0 ± 2.1
V8	46.42 ± 0.014	46.4 ± 1.5	46.4 ± 1.5	46.1 ± 1.5

Fig. 1. Phantom sample statistics of VPM, PGPM, and PERK T_1, T_2 estimates and NIST NMR reference measurements [52]. Plot markers and error bars indicate sample means and sample standard deviations computed over ROIs within the 14 vials labeled and color-coded in Fig. S.7. Yellow box boundaries indicate projections of the PERK sampling distribution's support $\text{supp}(p_{x,\nu})$. Missing markers lie outside axis limits. Corresponding tables replicate sample means $\pm$ sample standard deviations for vials within $\text{supp}(p_{x,\nu})$. Each value is rounded off to the highest place value of its (unreported) standard error, computed via formulas in [63]. 'V#' indicates vial numbers. All values are reported in milliseconds. Within $\text{supp}(p_{x,\nu})$, VPM, PGPM, and PERK estimates agree excellently with each other and reasonably with NMR measurements.

receive head array. To address bulk motion between scans, we rigidly registered coil-combined images to a reference before parameter estimation. All other data acquisition, image reconstruction, and parameter estimation details are the same as in phantom experiments (acquisition and reconstruction details are reported in [19]). VPM took 838s; PGPM took 2178s; and PERK training and testing respectively took 4.2s and 1.6s.

Fig. 2 compares VPM, PGPM, and PERK M_0, T_1, T_2 estimates. The PERK M_0 estimate appears smoothed (although no spatial regularization was used) but is otherwise very similar to the VPM and PGPM M_0 estimates. Narrow display ranges emphasize that VPM, PGPM, and PERK T_1, T_2 estimates discern cortical WM/GM boundaries similarly, though PERK T_1 estimates are noticeably highest in some WM regions. VPM, PGPM, and PERK T_2 estimates are nearly indistinguishable in lateral regions but disagree somewhat in medial regions close to cerebrospinal fluid (CSF). We neither expect nor observe reasonable PERK performance in voxels containing CSF.

Table II summarizes sample statistics of VPM, PGPM, and PERK T_1, T_2 estimates, computed over four separate WM ROIs containing 96, 69, 224, and 148 voxels and one pooled cortical anterior GM ROI containing 156 voxels. Overall, VPM, PGPM, and PERK T_1, T_2 estimates are comparable. T_1 estimates in GM and T_2 estimates in WM/GM do not differ significantly. PERK T_1 estimates are significantly higher than VPM and PGPM T_1 estimates in one WM ROI; however, all T_1 estimates are well within the range of typical literature measurements at 3T (see [64], [69]).

TABLE II

In vivo SAMPLE MEANS $\pm$ SAMPLE STANDARD DEVIATIONS OF VPM, PGPM, AND PERK T_1, T_2 ESTIMATES, COMPUTED OVER COLOR-CODED ROIS INDICATED IN FIG. 2. EACH VALUE IS ROUNDED OFF TO THE HIGHEST PLACE VALUE OF ITS (UNREPORTED) STANDARD ERROR, COMPUTED VIA FORMULAS IN [63]. ALL VALUES ARE IN MILLISECONDS.

	ROI	VPM	PGPM	PERK
T_1	AR WM	778 ± 28	779 ± 27	832 ± 31
	AL WM	731 ± 37	713 ± 33	725 ± 41
	PR WM	805 ± 52	796 ± 51	831 ± 51
	PL WM	789 ± 40	788 ± 38	815 ± 42
	A GM	1120 ± 180	1120 ± 180	$1150 \pm 170.$
T_2	AR WM	40.0 ± 1.29	40.0 ± 1.27	41.18 ± 0.94
	AL WM	39.7 ± 1.7	39.7 ± 1.7	41.3 ± 1.02
	PR WM	43.0 ± 2.7	43.0 ± 2.7	43.7 ± 2.6
	PL WM	43.0 ± 1.8	43.0 ± 1.8	43.5 ± 1.36
	A GM	53.5 ± 11.8	53.4 ± 11.7	53.3 ± 11.6

VII. DISCUSSION

The single-slice experiments show that PERK can achieve similar WM/GM T_1, T_2 estimation performance as

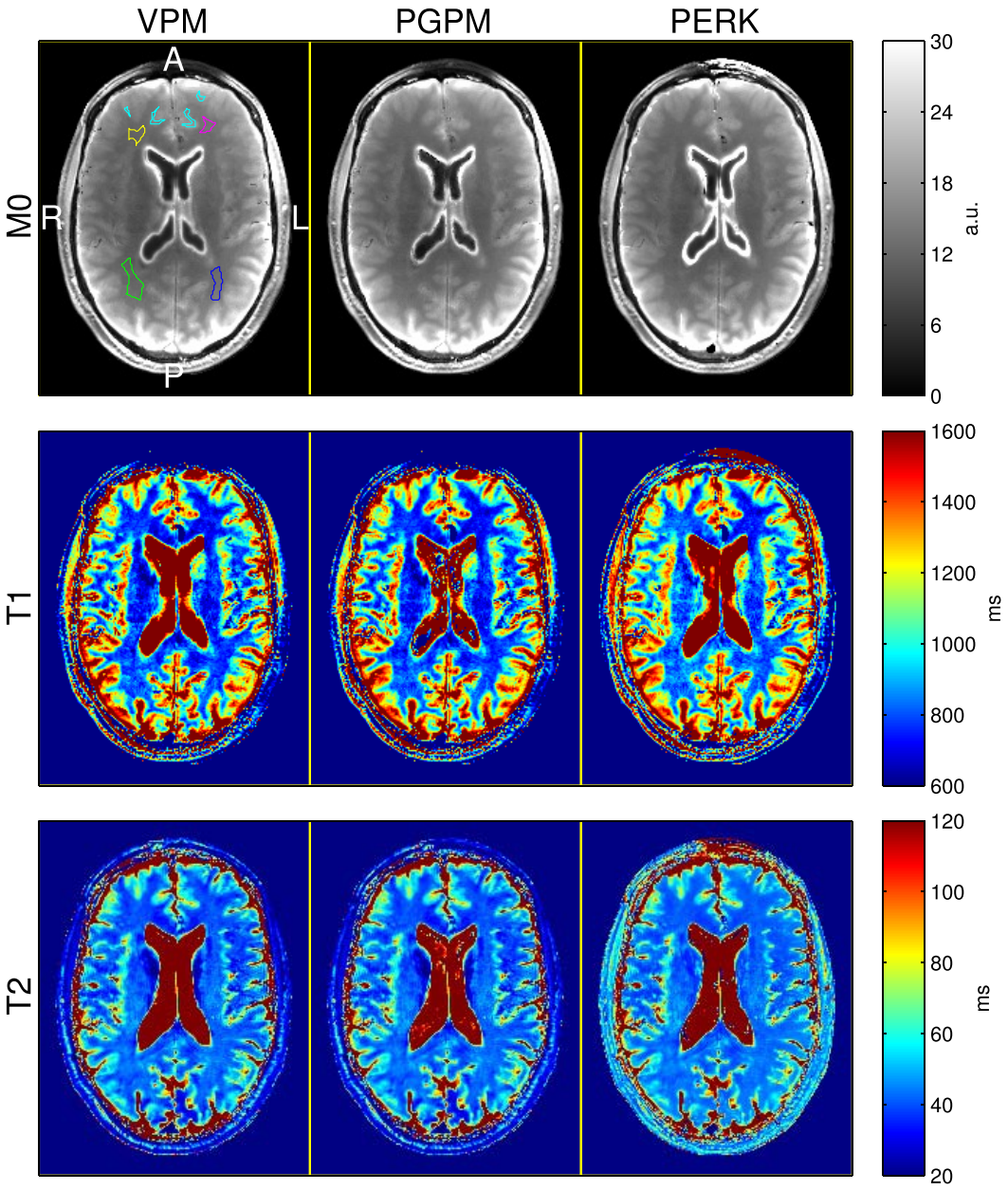


Fig. 2. VPM, PGPM and PERK estimates of M_0, T_1, T_2 in the brain of a healthy volunteer. Separate WM ROIs are distinguished by anterior/posterior (A/P) and right/left (R/L) directions. Four small anterior cortical GM polygons are pooled into a single GM ROI. Images are cropped in post-processing for display.

dictionary-based grid search via VPM or iterative optimization via PGPM, but in more than 2 orders of magnitude less time. This acceleration factor will grow to at least 3 orders of magnitude for T_1, T_2 estimation over a typical full imaging volume (because PERK training time scales negligibly with the number of voxels) and may grow even higher for full-volume parameter estimation in problems involving more unknowns per voxel (see [37] for a demonstration in simulation). Even with recent low-rank dictionary approximations [13], [17], [20], [21] dictionary-based methods are unlikely to achieve the large-scale speed of PERK.

PERK also handles known parameters $\boldsymbol{\nu}$ more naturally than does dictionary-based grid search. Grid search necessitates pre-clustering $\boldsymbol{\nu}$ voxel values and generating one dictionary per cluster; however, it is in general unclear *a priori* how many clusters are needed to balance accuracy and computation. In contrast, PERK simply considers the coordinates of each $\boldsymbol{\nu}$ sample as additional regressor dimensions. As the Gaussian PERK estimator is continuous in $\boldsymbol{\nu}$ (and $\boldsymbol{\alpha}$), Gaussian PERK does not suffer from either cluster (or grid) quantization bias.

Interestingly, PERK storage requirements grow more directly with regressor dimension P than with regressand dimension L. Using formulas for rank-one covariance matrix updates, constructing $\widehat{\mathbf{x}}(\cdot)$ element-wise via L evaluations of (23) can be implemented to use $O\left(Z^2\right)$ memory units when $\rho_l \leftarrow \rho \ \forall l \in \{1, \ldots, L\}$ (as recommended in §V.B.2). Direct

application of [47, Proposition 4] to the case of Gaussian kernel (12) reveals that Z should be scaled subquadratically but superlinearly with P to conservatively maintain a given threshold of maximal kernel approximation error. Thus, PERK memory requirements need grow no faster than $O\left(P^4\right)$ to maintain a given level of kernel approximation error.

The $O\left(P^4\right)$ PERK memory requirement ensures improvement over large-scale grid search in modestly overdetermined estimation problems, *i.e.* when $P \approx L$. In applications where the number of measurements far exceeds L (*e.g.*, MR fingerprinting [11]), PERK may still provide performance gains if images are projected [13] or directly reconstructed [20] into a low-dimensional measurement subspace prior to per-voxel processing. Using this idea, we recently applied PERK to MR fingerprinting in [70].

Phantom experiments most clearly demonstrate that while PERK T_1, T_2 estimates are accurate within a properly selected training range, PERK may extrapolate poorly outside the sampling distribution's support (an improperly selected support can significantly degrade performance; see §S.V.B for a demonstration). If more graceful degradation is desired, it may be helpful to additionally fit coefficients of a low-order polynomial and thereby form estimates of form, *e.g.*, $\widehat{x}_l(\mathbf{p}) := \widehat{h}_l(\mathbf{p}) + \widehat{b}_l + \widehat{\mathbf{c}}_l^{\mathsf{T}}\mathbf{p}$. However, greater model complexity may require more training samples to prevent overfitting.

In vivo experiments demonstrated that VPM, PGPM, and PERK T_1, T_2 estimates are overall comparable in WM and GM regions of interest. Nevertheless, small but consistently unidirectional discrepancies persist between the ML and PERK T_1 estimates in WM, one of which is statistically significant. These subtle discrepancies may indicate that ML and PERK estimators behave differently in regions with increased model mismatch. One possible source of *in vivo* model mismatch could be diffusive signal loss, to which DESS is especially sensitive [71], [72]. In particular, unaccounted diffusive signal loss could reduce the DESS second echo's already low SNR in WM to a point where non-Gaussian noise statistics become important to consider. Whereas PERK was trained with simulated data corrupted by Rician-distributed noise, the ML estimators used in this work take a (standard) Gaussian noise assumption and may thus be more prone than PERK to noise-related bias at low SNR. Taking these statements together, unaccounted diffusive effects might bias Gaussian ML estimators more than a properly trained PERK estimator and might explain minor discrepancies between ML and PERK T_1 estimates in WM.

The present formulation constructs separate scalar estimators for each coordinate of $\widehat{\mathbf{x}}$. A natural extension might instead seek to construct vector estimators that consist of linear combinations of vector features that reside in an RKHS of vector-valued functions (see [73] for a review). Here, the associated reproducing kernel would now be matrix-valued and might encode expected dependencies among the outputs of $\widehat{\mathbf{x}}$. With enough training points, the resulting vector estimator could achieve improved estimator performance in terms of accuracy and precision, at the expense of tuning more model parameters and increased computational burden.

In this work, we trained PERK using simulated training data corrupted by noise realizations drawn from a single noise distribution, whose statistics were estimated once from background regions of unlabeled test image data. This training strategy produced reasonable results perhaps in part because our experiments used fully-sampled Cartesian data, for which coil-combined images exhibit little spatial variation in the noise distribution due to receive coil sensitivity spatial variation [74]. To apply PERK in applications where input measurement images exhibit large spatial variation in the noise variance (*e.g.*, multiple-coil acquisitions with parallel imaging acceleration), it may be advantageous to train PERK using simulated training data corrupted by noise realizations drawn from an appropriate distribution over noise distributions. If noise variance maps are available, one could alternately train several PERK estimators with training datasets corrupted by different amounts of noise and apply each estimator to correspondingly noisy measurement image voxels.

Because there is ambiguity in MR data scale due to receive gains and other amplitude scaling factors, it is desirable to construct an estimator that is unaffected by changes in measurement scale between training and testing. In experiments, we address scaling ambiguity by setting the marginal M_0 sampling distribution p_{M_0} based on test measurements, thereby matching simulated training measurement scale to test measurement scale. This strategy would require retraining between acquisitions that are different in scale but are otherwise identical, which may be undesirable in practice. As an alternative, one could preprocess each noisy training regressor and each noisy test measurement by rescaling each such that (without loss of generality) its first entry is unity, is subsequently uninformative, and can thus be safely pruned to reduce problem dimensionality. Training and testing estimators (for latent parameters other than M_0) using these preprocessed regressors and test points is then largely invariant to the support of p_{M_0} [70]. One drawback to this approach is that normalization by noisy training regressors and test measurements could increase estimation variance.

As explained further in §V.B, we chose to train PERK after observation of unlabeled test data, a strategy that permits automatic selection of some tuning parameters but requires training at test time. Other applications may require many more training points than was required in our experiments for reasonable PERK performance, in which case such online training might be less practical. Using our PERK implementation, offline training would require additional selection of test measurement scale, known object parameter distribution p_ν, and noise variance σ^2. Test measurement scale selection could be avoided using the scale-invariant training strategy discussed in the previous paragraph. As emphasized in §V.B.1 and demonstrated in §S.V.B, PERK performance is quite sensitive to the object parameter distribution's support, and so at least the support of p_ν would need to be carefully selected based on separate prior parameter estimates or problem-specific intuition. As demonstrated in §S.III, PERK performs best when training and testing data noise statistics coincide but degrades gracefully with mild levels of mismatch, so σ^2 could be selected based on separate SNR approximations.

As an alternative to PERK, researchers have recently proposed MRI parameter estimation via deep neural network learning [75], [76]. Deep learning requires enormous numbers of training points to train many model parameters without overfitting, and its limited theoretical basis renders its practical use largely an art. Here, we have introduced and investigated PERK with an emphasis on its simplicity and its relatively intuitive model selection (see §V.B); a thorough comparison with deep learning is a possible topic for future work.

VIII. Conclusion

This paper has introduced PERK, a fast and general method for dictionary-free MRI parameter estimation. PERK first uses prior parameter/noise distributions and a general nonlinear MR signal model to simulate many parameter-measurement training points and then constructs a nonlinear regression function from these training points using linear combinations of nonlinear kernels. We have demonstrated PERK for T_1, T_2 estimation from optimized SPGR/DESS acquisitions [19], a simple application where it is straightforward to validate PERK estimates against gold-standard VPM estimates, iterative PGPM estimates, and NIST reference measurements. Numerical simulations showed that PERK achieves T_1, T_2 RMSE comparable to VPM and PGPM in WM- and GM-like voxels. Phantom experiments showed that within a properly chosen sampling distribution support, VPM, PGPM, and PERK estimates agree excellently with each other and reasonably with NIST NMR measurements. *In vivo* experiments showed that VPM, PGPM, and PERK produce comparable T_1 estimates and nearly indistinguishable T_2 estimates in WM and GM ROIs. PERK used identical model selection parameters across all simulations and experiments and consistently provided at least a 140× acceleration over VPM and PGPM. This acceleration factor may increase by several orders of magnitude for estimation problems involving more latent parameters per voxel [27], [37].

Acknowledgments

The authors thank Dr. Kathryn Keenan and Dr. Stephen Russek at NIST for generously lending a prototype [77] (used during acquisition testing) of the High Precision Devices® phantom. The authors also thank the reviewers for their helpful comments and suggestions.

References

[1] F. Bloch, "Nuclear induction," *Phys. Rev.*, vol. 70, nos. 7–8, pp. 460–474, Oct. 1946.

[2] H. C. Torrey, "Bloch equations with diffusion terms," *Phys. Rev.*, vol. 104, no. 3, pp. 563–565, 1956.

[3] H. M. McConnell, "Reaction rates by nuclear magnetic resonance," *J. Chem. Phys.*, vol. 28, no. 3, pp. 430–431, Mar. 1958.

[4] N. Bloembergen, E. M. Purcell, and R. V. Pound, "Relaxation effects in nuclear magnetic resonance absorption," *Phys. Rev.*, vol. 73, no. 7, pp. 679–712, Apr. 1948.

[5] D. Le Bihan *et al.*, "Diffusion tensor imaging: Concepts and applications," *J. Magn. Reson. Imag.*, vol. 13, no. 4, pp. 534–546, Apr. 2001.

[6] A. Mackay, K. Whittall, J. Adler, D. Li, D. Paty, and D. Graeb, "*In vivo* visualization of myelin water in brain by magnetic resonance," *Magn. Reson. Med.*, vol. 31, no. 6, pp. 673–677, Jun. 1994.

[7] J. P. Haldar, J. Anderson, and S. W. Sun, "Maximum likelihood estimation of T_1 relaxation parameters using VARPRO," in *Proc. Int. Soc. Mag. Res. Med.*, 2007, p. 41. [Online]. Available: http://cds.ismrm.org/ismrm-2007/files/00041.pdf

[8] D. Hernando, J. P. Haldar, B. P. Sutton, J. Ma, P. Kellman, and Z.-P. Liang, "Joint estimation of water/fat images and field inhomogeneity map," *Magn. Reson. Med.*, vol. 59, no. 3, pp. 571–580, Mar. 2008.

[9] J. K. Barral, E. Gudmundson, N. Stikov, M. Etezadi-Amoli, P. Stoica, and D. G. Nishimura, "A robust methodology for *in vivo* T_1 mapping," *Magn. Reson. Med.*, vol. 64, no. 4, pp. 1057–1067, Oct. 2010.

[10] E. Staroswiecki, K. L. Granlund, M. T. Alley, G. E. Gold, and B. A. Hargreaves, "Simultaneous estimation of T_2 and apparent diffusion coefficient in human articular cartilage *in vivo* with a modified three-dimensional double echo steady state (DESS) sequence at 3 T," *Magn. Reson. Med.*, vol. 67, no. 4, pp. 1086–1096, 2012.

[11] D. Ma *et al.*, "Magnetic resonance fingerprinting," *Nature*, vol. 495, pp. 187–193, Mar. 2013.

[12] J. D. Trzasko, P. M. Mostardi, S. J. Riederer, and A. Manduca, "Estimating T_1 from multichannel variable flip angle SPGR sequences," *Magn. Reson. Med.*, vol. 69, no. 6, pp. 1787–1794, 2013.

[13] D. F. McGivney *et al.*, "SVD compression for magnetic resonance fingerprinting in the time domain," *IEEE Trans. Med. Imag.*, vol. 33, no. 12, pp. 2311–2322, Dec. 2014.

[14] B. Zhao, F. Lam, and Z.-P. Liang, "Model-based MR parameter mapping with sparsity constraints: Parameter estimation and performance bounds," *IEEE Trans. Med. Imag.*, vol. 33, no. 9, pp. 1832–1844, Sep. 2014.

[15] N. Ben-Eliezer, D. K. Sodickson, and K. T. Block, "Rapid and accurate T_2 mapping from multi–spin-echo data using Bloch-simulation-based reconstruction," *Magn. Reson. Med.*, vol. 73, no. 2, pp. 809–817, Feb. 2015.

[16] B. Zhao, W. Lu, T. K. Hitchens, F. Lam, C. Ho, and Z.-P. Liang, "Accelerated MR parameter mapping with low-rank and sparsity constraints," *Magn. Reson. Med.*, vol. 74, no. 2, pp. 489–498, Aug. 2015.

[17] S. F. Cauley, W. Lu, T. K. Hitchens, F. Lam, C. Ho, and Z.-P. Liang, "Fast group matching for MR fingerprinting reconstruction," *Magn. Reson. Med.*, vol. 74, no. 2, pp. 523–528, Aug. 2015.

[18] B. Zhao, K. Setsompop, H. Ye, S. F. Cauley, and L. L. Wald, "Maximum likelihood reconstruction for magnetic resonance fingerprinting," *IEEE Trans. Med. Imag.*, vol. 35, no. 8, pp. 1812–1823, Aug. 2016.

[19] G. Nataraj, J.-F. Nielsen, and J. A. Fessler, "Optimizing MR scan design for model-based T_1, T_2 estimation from steady-state sequences," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 467–477, Feb. 2017.

[20] J. Assländer, M. A. Cloos, F. Knoll, D. K. Sodickson, J. Hennig, and R. Lattanzi, "Low rank alternating direction method of multipliers reconstruction for MR fingerprinting," *Magn. Reson. Med.*, vol. 79, no. 1, pp. 83–96, 2018.

[21] M. Yang *et al.*, "Low rank approximation methods for MR fingerprinting with large scale dictionaries," *Magn. Reson. Med.*, vol. 79, no. 4, pp. 2392–2400, 2018.

[22] D. A. Feinberg, L. E. Crooks, P. Sheldon, J. H. Iii, J. Watts, and M. Arakawa, "Magnetic resonance imaging the velocity vector components of fluid flow," *Magn. Reson. Med.*, vol. 2, no. 6, pp. 555–566, Dec. 1985.

[23] D. S. Tuch, V. J. Wedeen, A. M. Dale, J. S. George, and J. W. Belliveau, "Conductivity tensor mapping of the human brain using diffusion tensor MRI," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 20, pp. 11697–11701, Sep. 2001.

[24] K. Sekihara, S. Matsui, and H. Kohno, "NMR imaging for magnets with large nonuniformities," *IEEE Trans. Med. Imag.*, vol. MI-4, no. 4, pp. 193–199, Dec. 1985.

[25] G. R. Morrell, "A phase-sensitive method of flip angle mapping," *Magn. Reson. Med.*, vol. 60, no. 4, pp. 889–894, Oct. 2008.

[26] L. I. Sacolick, F. Wiesinger, I. Hancu, and M. W. Vogel, "B_1 mapping by Bloch-Siegert shift," *Magn. Reson. Med.*, vol. 63, no. 5, pp. 1315–1322, May 2010.

[27] G. Nataraj, J.-F. Nielsen, and J. A. Fessler, "Myelin water fraction estimation from optimized steady-state sequences using kernel ridge regression," in *Proc. Int. Soc. Mag. Res. Med.*, 2017, p. 5076.

[28] S. C. L. Deoni, B. K. Rutt, T. Arun, C. Pierpaoli, and D. K. Jones, "Gleaning multicomponent T_1 and T_2 information from steady-state imaging data," *Magn. Reson. Med.*, vol. 60, no. 6, pp. 1372–1387, Dec. 2008.

[29] S. C. L. Deoni, L. Matthews, and S. H. Kolind, "One component? Two components? Three? The effect of including a nonexchanging 'free' water component in multicomponent driven equilibrium single pulse observation of T_1 and T_2," *Magn. Reson. Med.*, vol. 70, no. 1, pp. 147–154, Jul. 2013.

[30] G. S. Kimeldorf and G. A. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *Ann. Math. Stat.*, vol. 41, no. 2, pp. 495–502, Apr. 1970.

[31] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, May 1950. [Online]. Available: http://www.jstor.org/stable/1990404

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[33] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *Proc. Int. Conf. Mach. Learn.*, 1998, pp. 515–521.

[34] C. Huang, C. G. Graff, E. W. Clarkson, A. Bilgin, and M. I. Altbach, "T_2 mapping from highly undersampled data by reconstruction of principal component coefficient maps using compressed sensing," *Mag. Res. Med.*, vol. 67, no. 5, pp. 1355–1366, May 2012.

[35] C. Huang, A. Bilgin, T. Barr, and M. I. Altbach, "T_2 relaxometry with indirect echo compensation from highly undersampled data," *Magn. Reson. Med.*, vol. 70, no. 4, pp. 1026–1037, Oct. 2013.

[36] L. Zhao, X. Feng, and C. H. Meyer, "Direct and accelerated parameter mapping using the unscented Kalman filter," *Magn. Reson. Med.*, vol. 75, no. 5, pp. 1989–1999, May 2016.

[37] G. Nataraj, J.-F. Nielsen, and J. A. Fessler, "Dictionary-free MRI parameter estimation via kernel ridge regression," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Apr. 2017, pp. 5–9.

[38] Y. Chang, D. Liang, and L. Ying, "Nonlinear GRAPPA: A kernel approach to parallel MRI reconstruction," *Magn. Reson. Med.*, vol. 68, no. 3, pp. 730–740, Sep. 2012.

[39] U. Nakarmi, Y. Wang, J. Lyu, D. Liang, and L. Ying, "A kernel-based low-rank (KLR) model for low-dimensional manifold recovery in highly accelerated dynamic MRI," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2297–2307, Nov. 2017.

[40] S. P. Awate and R. T. Whitaker, "Multiatlas segmentation as nonparametric regression," *IEEE Trans. Med. Imag.*, vol. 33, no. 9, pp. 1803–1817, Sep. 2014.

[41] S. P. Awate, R. M. Leahy, and A. A. Joshi, "Kernel methods for Riemannian analysis of robust descriptors of the cerebral cortex," in *Information Processing in Medical Imaging*. Cham, Switzerland: Springer, 2017, pp. 28–40.

[42] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. Int. Conf. Comput. Learn. Theory*, 2001, pp. 416–426.

[43] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970. [Online]. Available: http://www.jstor.org/stable/1267351

[44] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.

[45] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magn. Reson. Med.*, vol. 34, no. 6, pp. 910–914, Dec. 1995.

[46] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. NIPS*, 2007, pp. 1177–1184. [Online]. Available: https://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines

[47] D. J. Sutherland and J. Schneider, "On the error of random Fourier features," in *Proc. Int. Conf. Uncertainty AI*, 2015, pp. 1–18. [Online]. Available: http://arxiv.org/abs/1506.02785

[48] W. Zongmin, "Generalized Bochner's theorem for radial function," *Approximation Theory Appl.*, vol. 13, no. 3, pp. 47–57, 1997.

[49] M. A. Woodbury, "Inverting modified matrices," Statist. Res. Group, Princeton Univ., Princeton, NJ, USA, Tech. Rep. 42, 1950.

[50] V. N. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[51] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ, USA: Princeton Univ. Press, 1946. [Online]. Available: http://press.princeton.edu/titles/391.html

[52] K. E. Keenan *et al.*, "Multi-site, multi-vendor comparison of T_1 measurement using ISMRM/NIST system phantom," in *Proc. Int. Soc. Magn. Reson. Med.*, 2016, p. 3290.

[53] Y. Zur, M. L. Wood, and L. J. Neuringer, "Spoiling of transverse magnetization in steady-state sequences," *Magn. Reson. Med.*, vol. 21, no. 2, pp. 251–263, Oct. 1991.

[54] H. Bruder, H. Fischer, R. Graumann, and M. Deimling, "A new steady-state imaging sequence for simultaneous acquisition of two MR images with clearly different contrasts," *Magn. Reson. Med.*, vol. 7, no. 1, pp. 35–42, May 1988.

[55] M. M. Siddiqui, "Statistical inference for Rayleigh distributions," *J. Res. Nat. Bureau Standards, D*, vol. 68D, no. 9, pp. 1005–1010, Sep. 1964. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/jres/68D/jresv68Dn9p1005_A1b.pdf

[56] G. Golub and V. Pereyra, "Separable nonlinear least squares: The variable projection method and its applications," *Inverse Problems*, vol. 19, no. 2, pp. R1–R26, Feb. 2003.

[57] J. Gong and J. P. Hornak, "A fast T_1 algorithm," *Magn. Reson. Imag.*, vol. 10, no. 4, pp. 623–626, 1992.

[58] J. He, Q. Liu, A. G. Christodoulou, C. Ma, F. Lam, and Z.-P. Liang, "Accelerated high-dimensional MR imaging with sparse sampling using low-rank tensors," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2119–2129, Sep. 2016.

[59] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2007, pp. 1027–1035. [Online]. Available: http://dl.acm.org/citation.cfm?id=1283494

[60] J. B. Rosen, "The gradient projection method for nonlinear programming. Part I. Linear constraints," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 1, pp. 181–217, 1960.

[61] D. P. Bertsekas, "Projected Newton methods for optimization problems with simple constraints," *SIAM J. Control Optim.*, vol. 20, no. 2, pp. 221–246, 1982.

[62] R. K. Gupta, "A new look at the method of variable nutation angle for the measurement of spin-lattice relaxation times using Fourier transform NMR," *J. Magn. Reson.*, vol. 25, no. 1, pp. 231–235, Jan. 1977.

[63] S. Ahn and J. A. Fessler, "Standard errors of mean, variance, and standard deviation estimators," Commun. Signal Process. Lab., Dept. EECS, Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. 413, Jul. 2003. [Online]. Available: http://web.eecs.umich.edu/~fessler/papers/lists/files/tr/stderr.pdf

[64] J. P. Wansapura, S. K. Holland, R. S. Dunn, and W. S. Ball, "NMR relaxation times in the human brain at 3.0 tesla," *J. Magn. Reson. Imag.*, vol. 9, no. 4, pp. 531–538, Apr. 1999.

[65] D. L. Collins *et al.*, "Design and construction of a realistic digital brain phantom," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 463–468, Jun. 1998.

[66] J. Pauly, P. Le Roux, D. Nishimura, and A. Macovski, "Parameter relations for the Shinnar-Le Roux selective excitation pulse design algorithm (NMR imaging)," *IEEE Trans. Med. Imag.*, vol. 10, no. 1, pp. 53–65, Mar. 1991.

[67] L. Ying and J. Sheng, "Joint image reconstruction and sensitivity estimation in SENSE (JSENSE)," *Magn. Reson. Med.*, vol. 57, no. 6, pp. 1196–1202, Jun. 2007.

[68] H. Sun, W. A. Grissom, and J. A. Fessler, "Regularized estimation of Bloch-Siegert $|B_1^+|$ maps in MRI," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 3646–3650.

[69] G. J. Stanisz *et al.*, "T_1, T_2 relaxation and magnetization transfer in tissue at 3 T," *Magn. Reson. Med.*, vol. 54, no. 3, pp. 507–512, Sep. 2005.

[70] G. Nataraj, M. Gao, J. Assländer, C. Scott, and J. A. Fessler, "Shallow learning with kernels for dictionary-free magnetic resonance fingerprinting," in *Proc. ISMRM Workshop MR Fingerprinting*, 2017, pp. 2–4.

[71] E. X. Wu and R. B. Buxton, "Effect of diffusion on the steady-state magnetization with pulsed field gradients," *J. Mag. Res.*, vol. 90, no. 2, pp. 243–253, Nov. 1990.

[72] C. E. Carney, S. T. S. Wong, and S. Patz, "Analytical solution and verification of diffusion effect in SSFP," *Magn. Reson. Med.*, vol. 19, no. 2, pp. 240–246, Jun. 1991.

[73] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2011-033, Jun. 2011. [Online]. Available: http://cbcl.mit.edu/publications/ps/MIT-CSAIL-TR-2011-033.pdf

[74] S. Aja-Fernández and G. Vegas-Sánchez-Ferrero, *Statistical Analysis of Noise in MRI: Modeling, Filtering and Estimation*. Cham, Switzerland: Springer, 2016.

[75] O. Cohen, B. Zhu, and M. S. Rosen. (2017). "Deep learning for rapid sparse MR fingerprinting reconstruction." [Online]. Available: http://arxiv.org/abs/1710.05267

[76] P. Virtue, S. X. Yu, and M. Lustig, "Better than real: Complex-valued neural nets for MRI fingerprinting," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 3953–3957.

[77] S. E. Russek *et al.*, "Characterization of NIST/ISMRM MRI system phantom," in *Proc. Int. Soc. Magn. Reson. Med.*, 2012, p. 2456. [Online]. Available: http://dev.ismrm.org//2012/2456.html

Supplementary Material for Dictionary-Free MRI PERK: Parameter Estimation via Regression with Kernels

Gopal Nataraj*, Jon-Fredrik Nielsen†, Clayton Scott*, and Jeffrey A. Fessler*

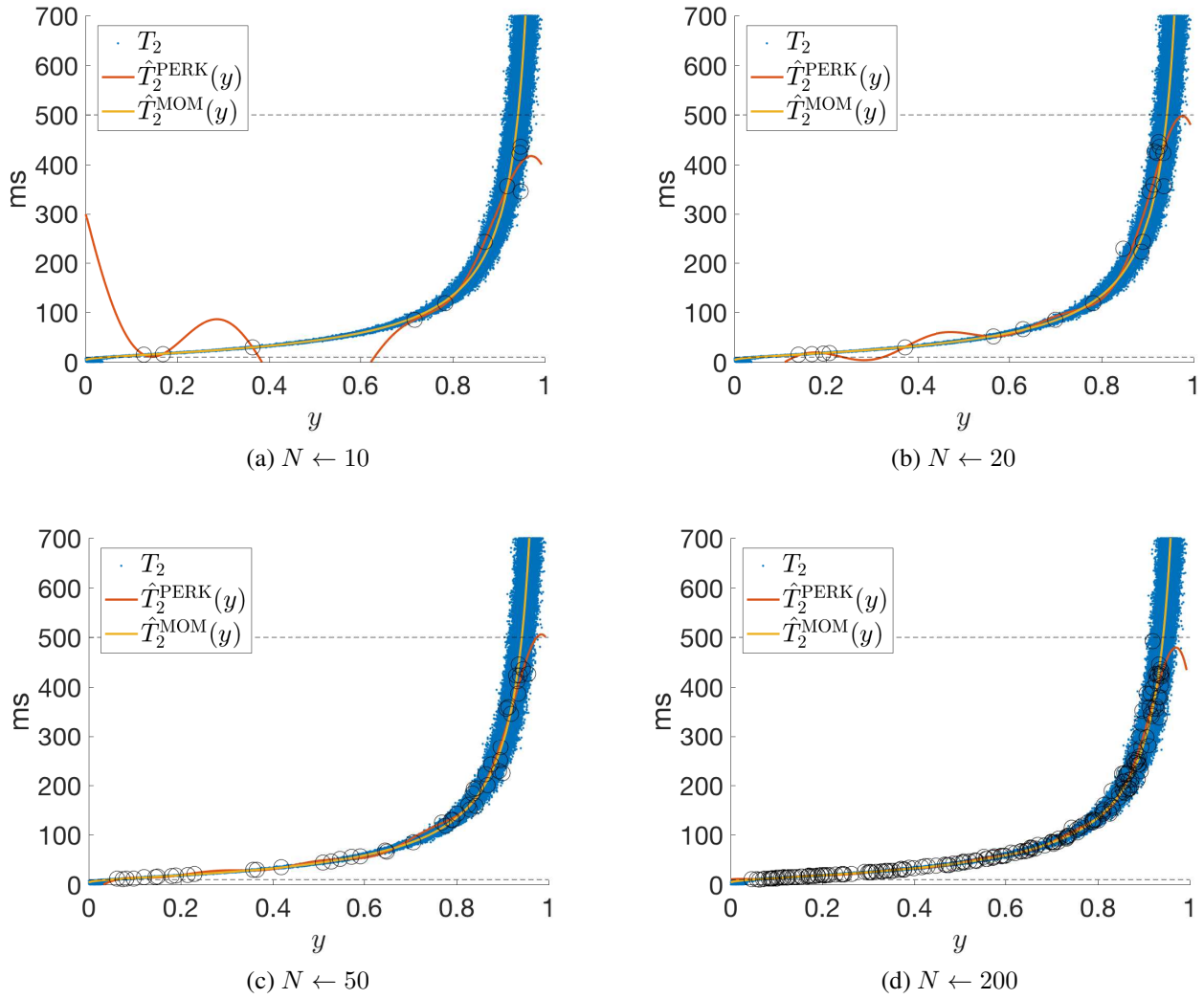
*Dept. of Electrical Engineering and Computer Science, University of Michigan
†Dept. of Biomedical Engineering, University of Michigan

This supplement provides further intuition, elaborates upon methodology details, and presents additional figures that could not be included in the main body of the manuscript [1] due to page restrictions. §S.I demonstrates and illustrates PERK in a simple toy problem. §S.II details our procedure for selecting free tuning parameters. §S.III investigates PERK performance sensitivity to mismatch between training and testing data noise variance. §S.IV presents estimated parameter images corresponding to numerical simulations presented in §VI.B. §S.V provides additional phantom results and discusses PERK performance degradation when trained with latent parameter distributions that have wider support than the parameter ranges used for optimizing the scan design in [2].

S.I Demonstration in a 1-D Toy Problem

To build intuition and for ease of visualization, we applied PERK to a simple one-dimensional toy problem, namely T_2 estimation from a single spin-echo measurement. We generated training data using a mono-exponential (unity-M_0) signal model $y = e^{-T_\mathrm{E}/T_2} + \epsilon$, where y is a complex spin-echo measurement, $T_\mathrm{E} \leftarrow 30$ms is the echo time and $\epsilon \sim \mathbb{C}\mathcal{N}(0, 0.01^2)$ is complex Gaussian noise. We sampled $N \leftarrow 10$, 20, 50, 200 regressands from T_2 sampling distribution $\mathsf{p}_{T_2} \leftarrow \mathrm{logunif}(10, 500)$ and took the magnitude of noisy complex signal model evaluations to generate corresponding magnitude regressors. We trained PERK separately using each of the four labeled training datasets, holding fixed hyperparameters $(\lambda, \rho) \leftarrow (2^{-1.5}, 2^{-20})$ that were manually chosen to aid in illustrating PERK's typical behavior.

Fig. S.1 illustrates the 1-D PERK estimator $\widehat{T}_2^{\mathrm{PERK}}$ and shows how its performance improves as N is increased. To produce each subfigure, we uniformly sampled 100,000 true (latent) T_2 values, evaluated the noisy signal model as in training to generate magnitude test points (blue dots), and evaluated each PERK estimator at the unlabeled test points (orange curves). For comparison, subfigures within Fig. S.1 also plot the intuitive method-of-moments (MOM) estimator $\widehat{T}_2^{\mathrm{MOM}}(\cdot) := -T_\mathrm{E}/\log|\cdot|$ (yellow curves). As N increases, $\widehat{T}_2^{\mathrm{PERK}}$ appears more similar to $\widehat{T}_2^{\mathrm{MOM}}$ within well-sampled regions of $\mathsf{supp}(\mathsf{p}_{T_2})$ (marked by dashed black lines). PERK will be more useful in nonlinear estimation problems where such a minimally biased and low-dimensional MOM estimator is unavailable.



(a) $N \leftarrow 10$ (b) $N \leftarrow 20$

(c) $N \leftarrow 50$ (d) $N \leftarrow 200$

Figure S.1: Illustrations of PERK for T_2 estimation from a single spin echo measurement. Subfigures vary the number N of PERK training points, marked with black circles. The orange and yellow curves plot PERK $\widehat{T}_2^{\mathrm{PERK}}$ and MOM $\widehat{T}_2^{\mathrm{MOM}}$ estimators evaluated at test points, marked with blue dots. Dashed black lines denote the sampling distribution support $\mathsf{supp}(\mathsf{p}_{T_2})$ over which each PERK estimator was trained. As N increases, $\widehat{T}_2^{\mathrm{PERK}}$ appears more similar to $\widehat{T}_2^{\mathrm{MOM}}$ within well-sampled regions of $\mathsf{supp}(\mathsf{p}_{T_2})$.

S.II Tuning Parameter Selection via Holdout

We selected Gaussian kernel bandwidth scaling parameter λ and regularization parameter ρ using the following offline holdout procedure in simulation. We discretized (λ, ρ) over a finely spaced grid spanning many orders of magnitude. As described in §VI.A, we trained a PERK estimator $\widehat{\mathbf{x}}_{\lambda,\rho}$ for each candidate model parameter setting. We evaluated each PERK estimator on a separate simulated dataset consisting of many samples from the training prior distribution $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}}$. We selected model parameters by exhaustively seeking a minimizer $\left(\widehat{\lambda}, \widehat{\rho}\right)$ of the "holdout" cost function

$$\Psi(\lambda, \rho) := \sqrt{\frac{1}{T}\sum_{t=1}^{T} \left\| [\mathrm{diag}(\mathbf{x}_t)]^{-1} (\widehat{\mathbf{x}}_{\lambda,\rho}(\mathbf{p}_t) - \mathbf{x}_t) \right\|_{\mathbf{W}}^2} \tag{S.1}$$

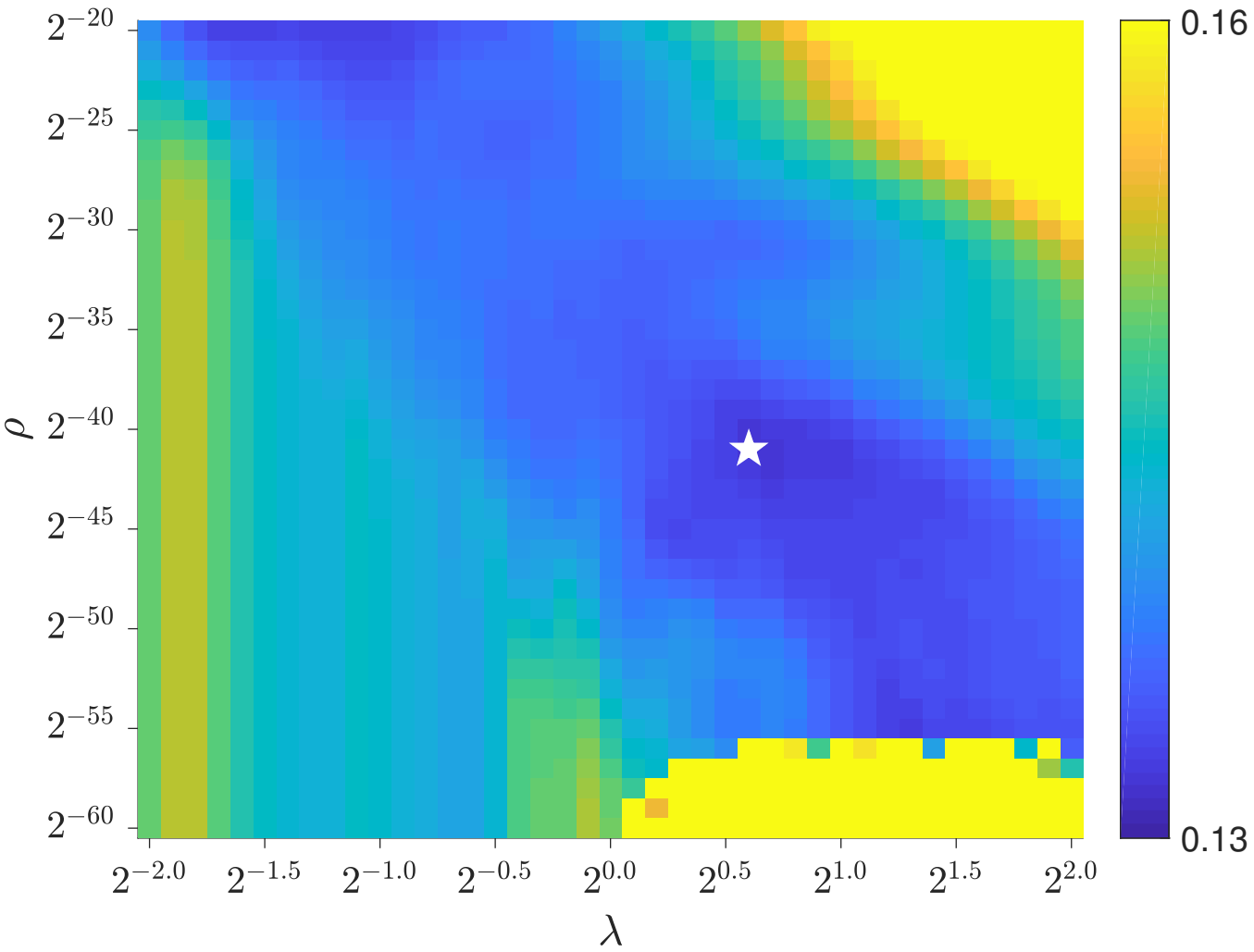


Figure S.2: Holdout criterion $\Psi(\lambda, \rho)$ versus Gaussian kernel bandwidth scaling parameter λ and regularization parameter ρ. Each pixel is the weighted normalized root mean squared error of a candidate PERK estimator, where the empirical mean over 10^5 test points approximates an expectation with respect to training prior distribution $\mathsf{p}_{\mathbf{x},\nu}$ and the weighting places emphasis on good T_1, T_2 estimation performance. A white star marks the minimizer $\left(\widehat{\lambda}, \widehat{\rho}\right) \leftarrow (2^{0.6}, 2^{-41})$.

where $t \in \{1, \ldots, T\}$ indexes T test points; each $\mathbf{x}_t$ is the true latent parameter corresponding to holdout test data point $\mathbf{p}_t$; and $\mathbf{W}$ is a diagonal unit-trace weighting matrix. Intuitively, $\Psi(\lambda, \rho)$ is the weighted normalized root mean squared error of PERK estimator $\widehat{\mathbf{x}}_{\lambda,\rho}$, where the mean approximates an expectation with respect to $\mathsf{p}_{\mathbf{x},\nu}$ and the latent parameter weighting is specified by $\mathbf{W}$.

Fig. S.2 plots $\Psi(\lambda, \rho)$ for $T \leftarrow 10^5$ test points and $\mathbf{W} \leftarrow \mathrm{diag}\left([0, 0.5, 0.5]^{\mathsf{T}}\right)$ selected to place equal emphasis on T_1, T_2 estimation. We chose our fine grid search range using a preliminary coarse grid search spanning a much wider range of (λ, ρ) values. Overall, we observe a broad range of (λ, ρ) values that yield similar cost function values. Holdout cost $\Psi(\lambda, \rho)$ gracefully increases with larger (λ, ρ) values due to under-fitting. For very small ρ values, $\Psi(\lambda, \rho)$ can be large because poorly conditioned matrix inversions cause machine imprecision to dominate estimation error. In all simulations and experiments, we fixed free model parameters to the minimizer $\left(\widehat{\lambda}, \widehat{\rho}\right) \leftarrow (2^{0.6}, 2^{-41})$, indicated by a white star.

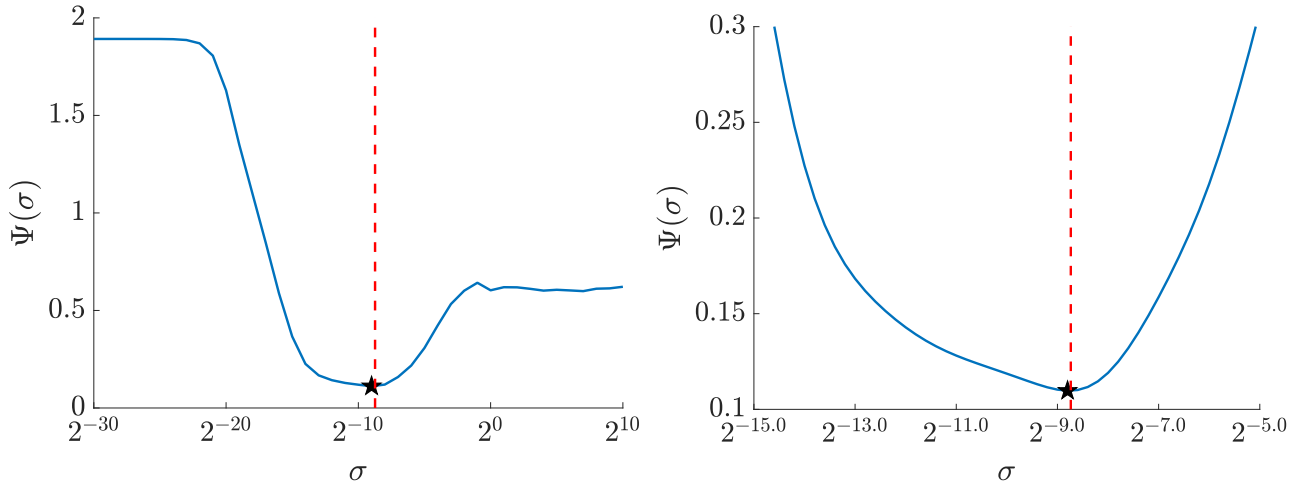


Figure S.3: Performance criterion $\Psi(\sigma)$ versus PERK training noise standard deviation σ, over two different scales. Similar to Fig. S.2, each point on the blue curve is the weighted normalized root mean squared error of a separately trained PERK estimator. In each subplot, a black star marks the performance criterion minimizer $\widehat{\sigma} \leftarrow 2^{-8.8}$ while a dashed red line marks the (latent) test data noise standard deviation $\sigma^\star \leftarrow 2^{-8.736}$. To within quantization error, PERK performs best when trained with training data whose noise statistics match test data noise statistics.

S.III Performance Sensitivity to Training Noise Variance

To assess the importance of training PERK with appropriately noisy training data, we investigated PERK's performance sensitivity to the standard deviation σ of the noise distribution from which noise realizations are drawn to generate training data. Instead of setting σ online as in other experiments, here we fixed σ offline to one of many discretized values spanning many orders of magnitude. Otherwise as described in §VI.A, we trained a PERK estimator for each σ setting. Similar to §S.II, we tested each PERK estimator on a separate simulated dataset consisting of 10^5 samples from training prior distribution $\mathsf{p}_{\mathbf{x},\nu}$. We assessed performance sensitivity by comparing evaluations $\Psi(\sigma)$ of holdout cost (S.1) at each σ setting.

Fig. S.3 plots $\Psi(\sigma)$ as σ is varied over two different scales. In each subplot, a black star marks the minimizer $\hat{\sigma} \leftarrow 2^{-8.8}$ while a dashed red line marks the (latent) test data noise standard deviation $\sigma^\star \leftarrow 2^{-8.736}$. To within quantization error, PERK performs best when trained with training data whose noise statistics match test data noise statistics. As measured by Ψ, PERK performance degrades by at most 10% for choices of $\sigma \in [2^{-10.2}, 2^{-8}]$. Results suggest that for good PERK performance, it is desirable to set σ to within about a factor of two of the test data noise standard deviation $\sigma^\star$. Nevertheless, there is a zone near $\sigma^\star$ where PERK performance is reasonably similar, indicating that PERK is somewhat robust to some misspecification of the noise level.

S.IV Numerical Simulations

Figs. S.4, S.5, and S.6 compare VPM, PGPM, and PERK estimates of M_0, T_1, T_2 respectively, alongside $10\times$ magnified absolute difference images with respect to the ground truth. Voxels not assigned WM- or GM-like relaxation times are masked out in post-processing for display. Table S.1 extends Table I to present M_0 in addition to T_1, T_2 sample statistics within WM- and GM-like ROIs. Difference images demonstrate that within WM- and GM-like voxels, all three methods exhibit low estimation error.

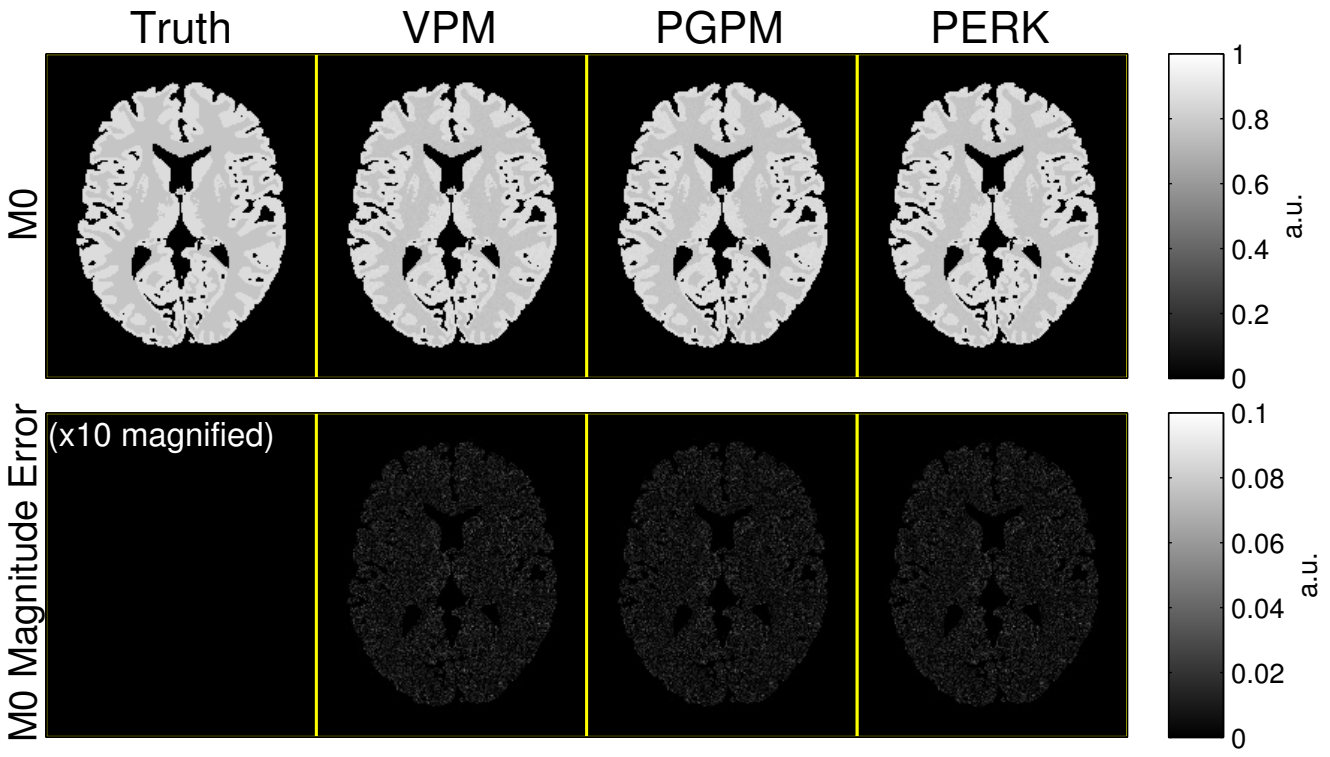


Figure S.4: M_0 VPM, PGPM, and PERK estimates and corresponding error images, in simulation. Magnitude error images are $10\times$ magnified. Voxels not assigned WM- or GM-like relaxation times are masked out in post-processing for display. Difference images demonstrate that all three M_0 estimates exhibit low estimation error. Table S.1 presents corresponding sample statistics.

	Truth	VPM	PGPM	PERK
WM M_0	0.77	0.7700 ± 0.00919 (0.0092)	0.76999 ± 0.00871 (0.00871)	0.77002 ± 0.00873 (0.00873)
GM M_0	0.86	0.8601 ± 0.01192 (0.0119)	0.8600 ± 0.01142 (0.0114)	0.8613 ± 0.01147 (0.0133)
WM T_1	832	832.1 ± 17.2 (17.2)	832.1 ± 16.2 (16.2)	833.0 ± 16.5 (16.5)
GM T_1	1331	1331.5 ± 31.1 (31.1)	1331.2 ± 29.7 (29.7)	1332.1 ± 30.4 (30.4)
WM T_2	79.6	79.61 ± 0.988 (0.988)	79.60 ± 0.952 (0.952)	79.46 ± 0.978 (0.989)
GM T_2	110.	110.02 ± 1.40 (1.40)	110.02 ± 1.35 (1.35)	109.91 ± 1.35 (1.35)

Table S.1: Sample means $\pm$ sample standard deviations (RMSEs) of VPM, PGPM, and PERK M_0, T_1, T_2 estimates, computed in simulation over 7810 WM-like and 9162 GM-like voxels. Each sample statistic is rounded off to the highest place value of its (unreported) standard error, computed via formulas in [3]. M_0 values are unitless. T_1, T_2 values are reported in milliseconds and were also reported in Table I.

S.V Phantom Experiments

S.V.A Training over a conservative sampling distribution support

Fig. S.7 compares VPM, PGPM, and PERK M_0, T_1, T_2 estimates in a quantitative phantom. Vials are enumerated in descending T_1, T_2 order. Vials whose T_1, T_2 values are within sampling distribution support $\mathsf{supp}(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}})$ (as measured by NIST NMR reference measurements [4]) have labels highlighted with yellow numbers. Here, $\mathsf{supp}(\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}})$ was chosen to reflect the ranges of latent parameter values for which the

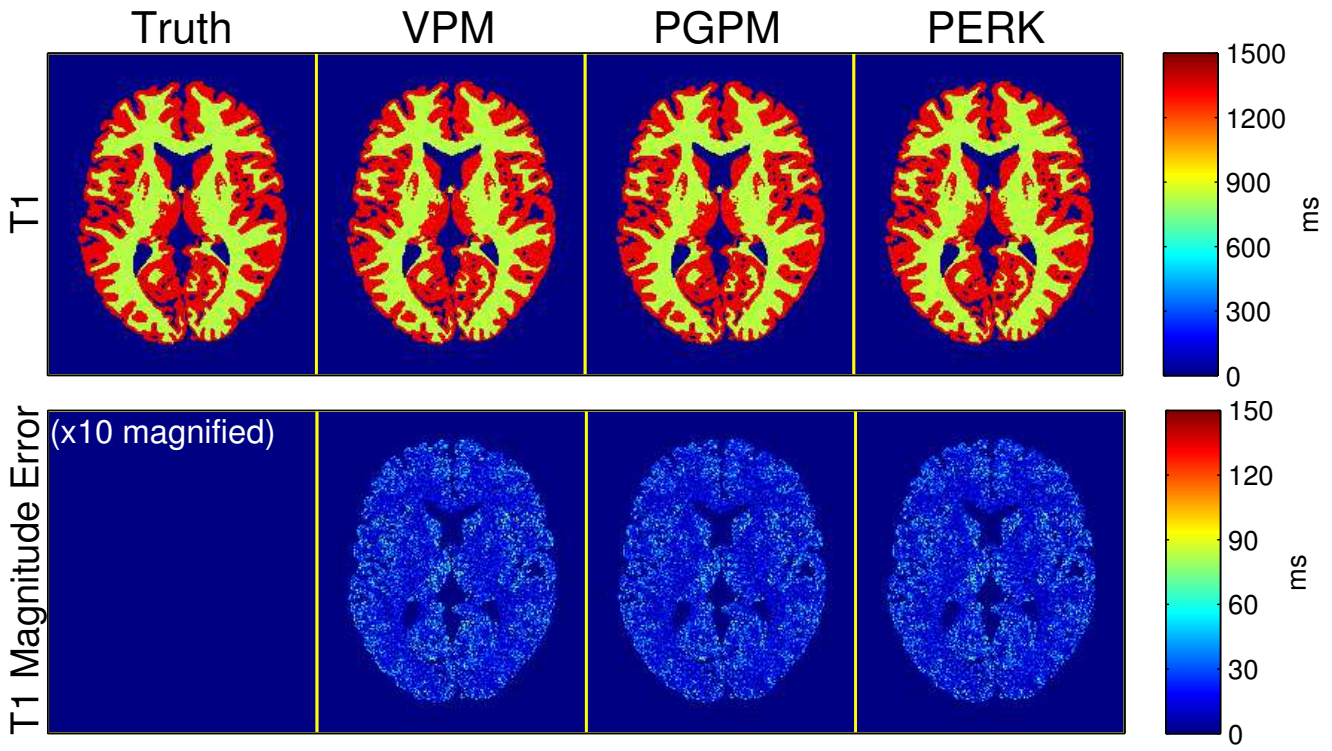


Figure S.5: T_1 VPM, PGPM, and PERK estimates and corresponding error images, in simulation. Magnitude error images are $10\times$ magnified. Voxels not assigned WM- or GM-like relaxation times are masked out in post-processing for display. Difference images demonstrate that all three T_1 estimates exhibit low estimation error. Tables I and S.1 both present the same corresponding sample statistics.

SPGR/DESS scan parameters were optimized in [2]. Circular ROIs are selected well away from vial encasings and correspond with sample statistics presented in Fig. 1. Distilled water surrounds the encased vials. Within the highlighted vials of interest, VPM, PGPM, and PERK estimates appear visually similar.

S.V.B Training over an aggressive sampling distribution support

Although the SPGR/DESS acquisition was optimized in [2] for a certain range of T_1, T_2 values, it is interesting to investigate how well PERK can perform outside that parameter range if presented (simulated) training data over a wider range of latent parameters. It is also interesting to explore whether using such a wider range of latent parameters for training degrades performance for the parameter range of primary interest. Thus, we repeated the phantom experiment described in §VI.C except now using a PERK estimator trained using a sampling prior distribution with broader support. We still assume a separable prior distribution $\mathsf{p}_{\mathbf{x},\boldsymbol{\nu}} \leftarrow \mathsf{p}_{M_0}\mathsf{p}_{T_1}\mathsf{p}_{T_2}\mathsf{p}_{\kappa}$ (with p_{M_0} and p_{κ} set as before) but now set $\mathsf{p}_{T_1} \leftarrow \text{logunif}(10^{1.5}, 10^{3.5})$ and $\mathsf{p}_{T_2} \leftarrow \text{logunif}(10^{0.5}, 10^{3.5})$ to have wider supports. These support endpoints now match the grid search support used by the VPM. All other training and testing details are unchanged from before.

Fig. S.8 is analogous to Fig. 1 in that it plots sample means and sample standard deviations computed within ROIs of VPM, PGPM, and PERK T_1, T_2 estimates, except now using a PERK estimator trained over the broader sampling distribution. Fig. S.9 presents corresponding images. The yellow boxes are unchanged from Fig. 1 and so their boundaries no longer correspond to projections of the PERK sampling distribution's support. Rather, they serve to clearly highlight that PERK estimator performance can significantly deteriorate even over the parameter range of interest, when trained using a range of parameters

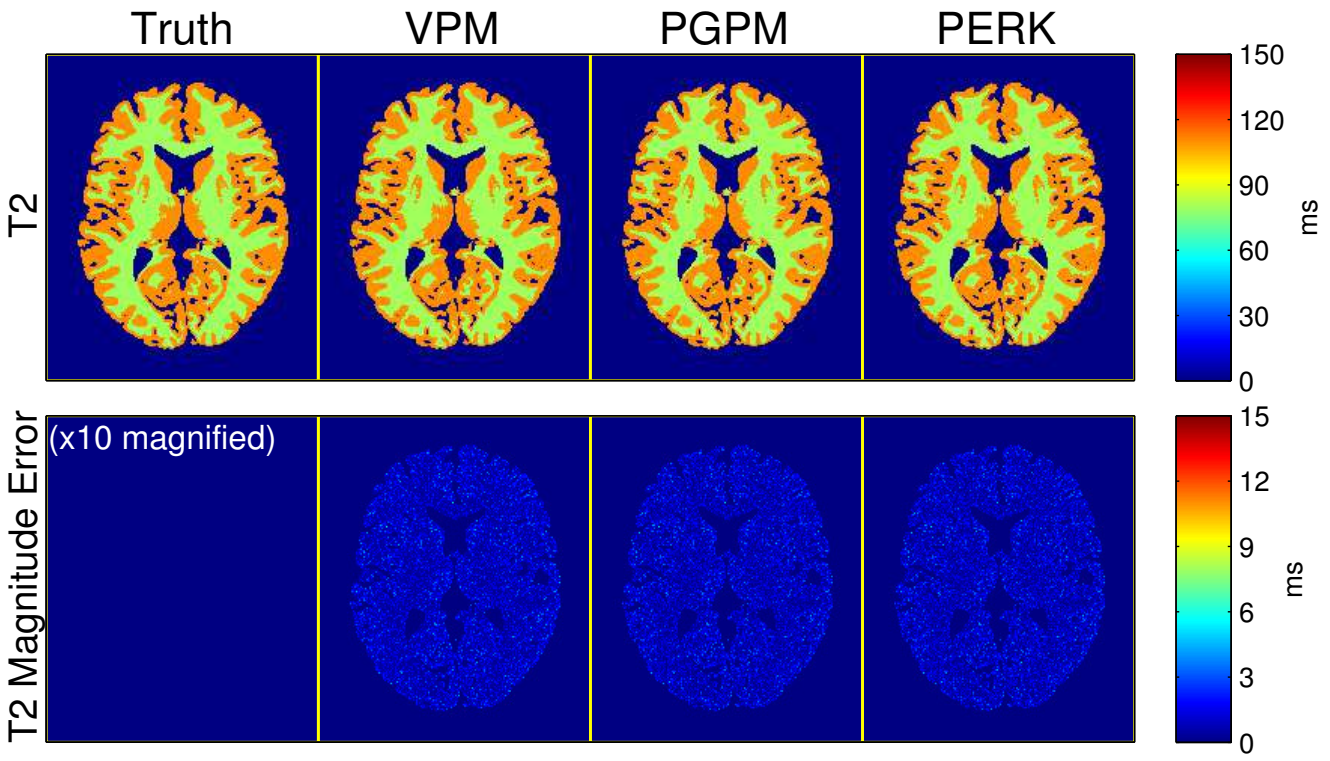


Figure S.6: T_2 VPM, PGPM, and PERK estimates and corresponding error images, in simulation. Magnitude error images are $10\times$ magnified. Voxels not assigned WM- or GM-like relaxation times are masked out in post-processing for display. Difference images demonstrate that all three T_1 estimates exhibit low estimation error. Tables I and S.1 both present the same corresponding sample statistics.

that exceeds the design criteria of the acquisition.

Fig. S.8 also tabulates sample means and sample standard deviations computed within ROIs of vials 4-8. Comparing again with Fig. 1, PERK T_2 estimation accuracy is more severely affected than T_1 estimation accuracy (interestingly, T_1 estimation accuracy is in fact improved for many vials). PERK T_1, T_2 estimation precision is consistently worse in vials 4-8 when trained over the broader sampling range.

These observations highlight the importance of considering acquisition design and parameter estimation in tandem, and with consideration of the latent parameter ranges of interest in a given application.

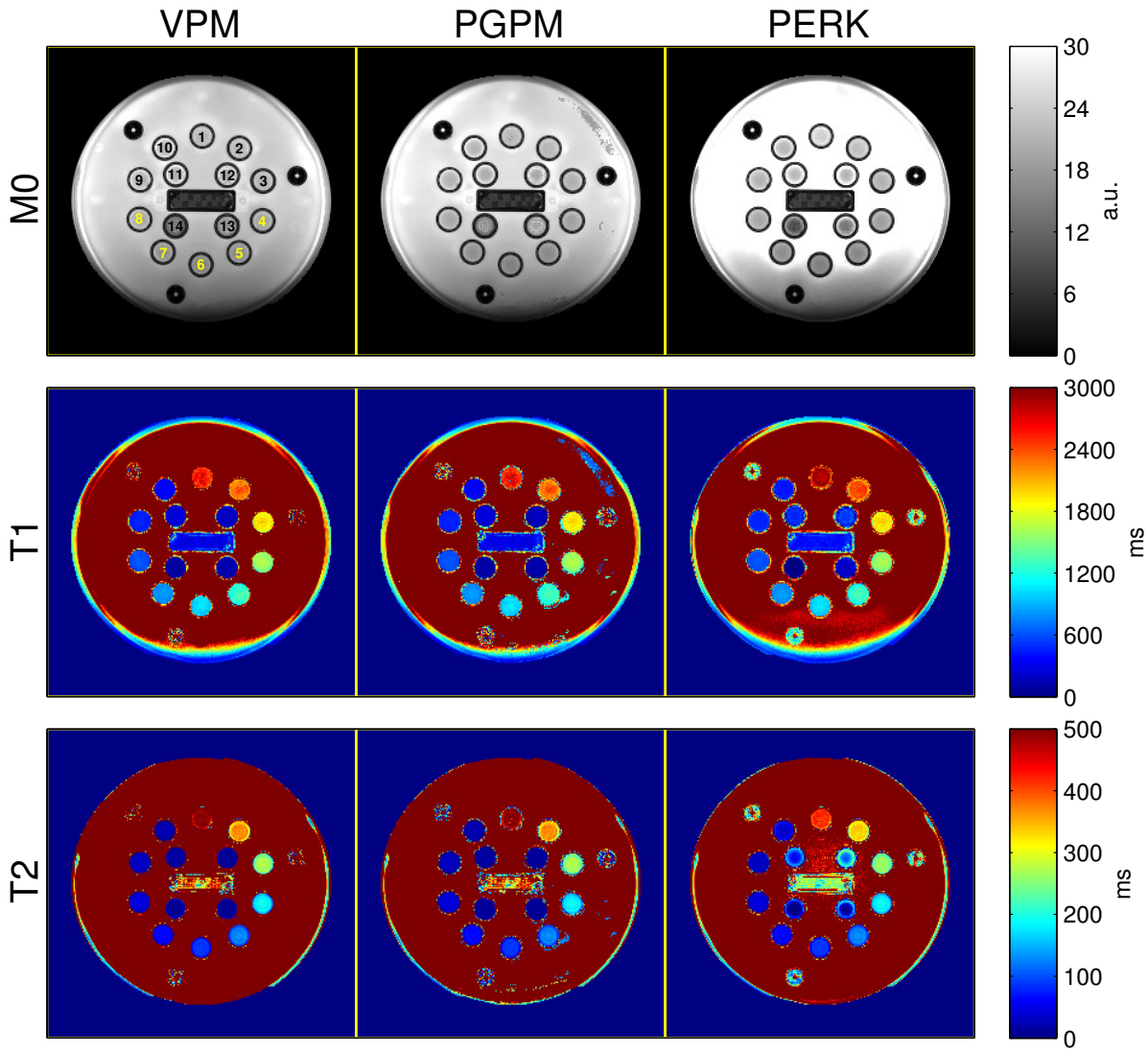


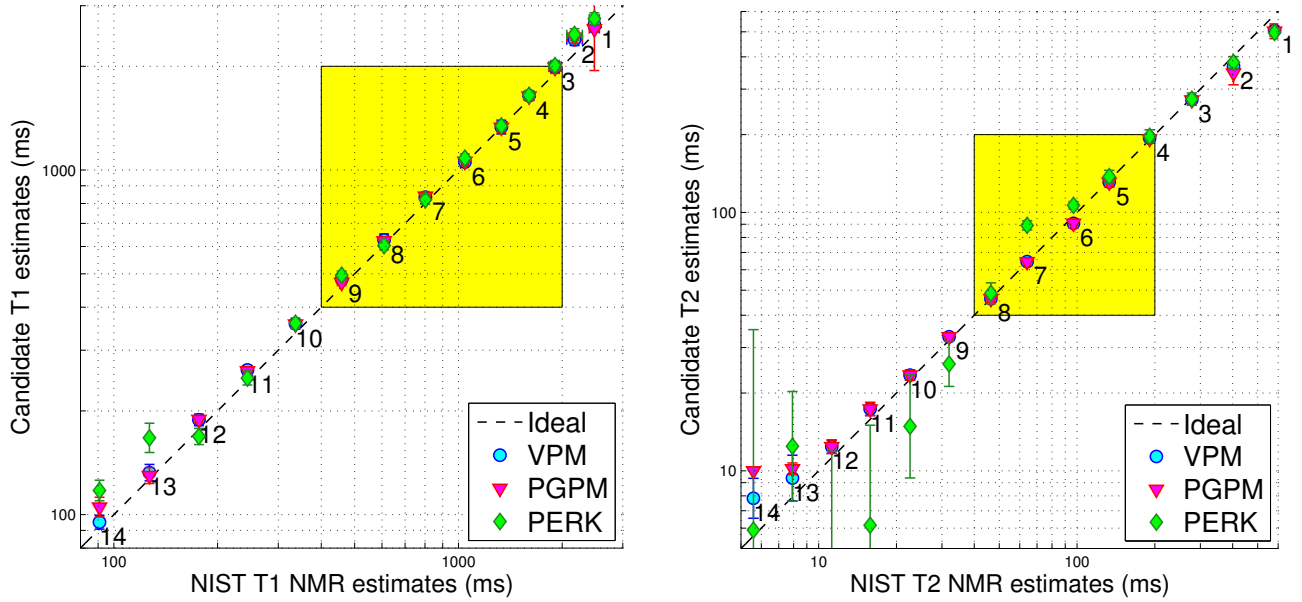
Figure S.7: VPM, PGPM, and PERK M_0, T_1, T_2 estimates in a quantitative phantom. Vials are enumerated and highlighted to correspond with markers and colored boxes in Fig. 1. PERK has only been trained to accurately estimate within vials 4-8; within these vials, VPM, PGPM, and PERK estimates appear visually similar.

References

[1] G. Nataraj, J.-F. Nielsen, C. Scott, and J. A. Fessler, "Dictionary-free MRI PERK: Parameter estimation via regression with kernels," *IEEE Trans. Med. Imag.*, 2018, to appear.

[2] G. Nataraj, J.-F. Nielsen, and J. A. Fessler, "Optimizing MR scan design for model-based T1, T2 estimation from steady-state sequences," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 467–77, Feb. 2017.

[3] S. Ahn and J. A. Fessler, "Standard errors of mean, variance, and standard deviation estimators," Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Tech. Rep. 413, Jul. 2003. [Online]. Available: http://web.eecs.umich.edu/~fessler/papers/lists/files/tr/stderr.pdf



	NMR	VPM	PGPM	PERK
V4 T_1	1604 ± 7.2	1645 ± 48	1639 ± 48	1649 ± 51
V5 T_1	1332 ± 0.8	1335 ± 61	1331 ± 41	$1343 \pm 40.$
V6 T_1	1044 ± 3.2	1055 ± 28	$1060. \pm 29$	1083 ± 32
V7 T_1	801.7 ± 1.70	834 ± 21	$840. \pm 23$	821 ± 25
V8 T_1	608.6 ± 1.03	627 ± 25	623 ± 12	604 ± 18
V4 T_2	190.94 ± 0.011	194 ± 5.5	193.1 ± 5.2	197 ± 11
V5 T_2	133.27 ± 0.073	131.2 ± 5.3	131 ± 5.5	138 ± 8
V6 T_2	96.89 ± 0.049	90.8 ± 3.5	90.8 ± 3.5	106.6 ± 3.6
V7 T_2	64.07 ± 0.034	64.6 ± 2.2	64.5 ± 2.1	89.2 ± 3.7
V8 T_2	46.42 ± 0.014	46.4 ± 1.5	46.4 ± 1.5	48.9 ± 4.6

Figure S.8: Phantom sample statistics of more aggressively trained VPM, PGPM, and PERK T_1, T_2 estimates and NIST NMR reference measurements [4]. Unlike analogous results in Fig. 1, here the PERK estimator was trained with a sampling distribution whose support extended well beyond the range of T_1, T_2 values for which the acquisition was optimized in [2]. Comparing to Fig. 1, we find that PERK estimator performance degrades within the highlighted T_1, T_2 range of interest. Plot markers and error bars indicate sample means and sample standard deviations computed over ROIs within the 14 vials labeled and color-coded in Fig. S.9. Corresponding tables replicate sample means $\pm$ sample standard deviations for vials within the highlighted range. Each value is rounded off to the highest place value of its (unreported) standard error, computed via formulas in [3]. All values are in milliseconds.

[4] K. E. Keenan, K. F. Stupic, M. A. Boss, S. E. Russek, T. L. Chenevert, P. V. Prasad, W. E. Reddick, K. M. Cecil, J. Zheng, P. Hu, and E. F. Jackson, "Multi-site, multi-vendor comparison of T1 measurement using ISMRM/NIST system phantom," in *Proc. Intl. Soc. Mag. Res. Med.*, 2016, p. 3290.

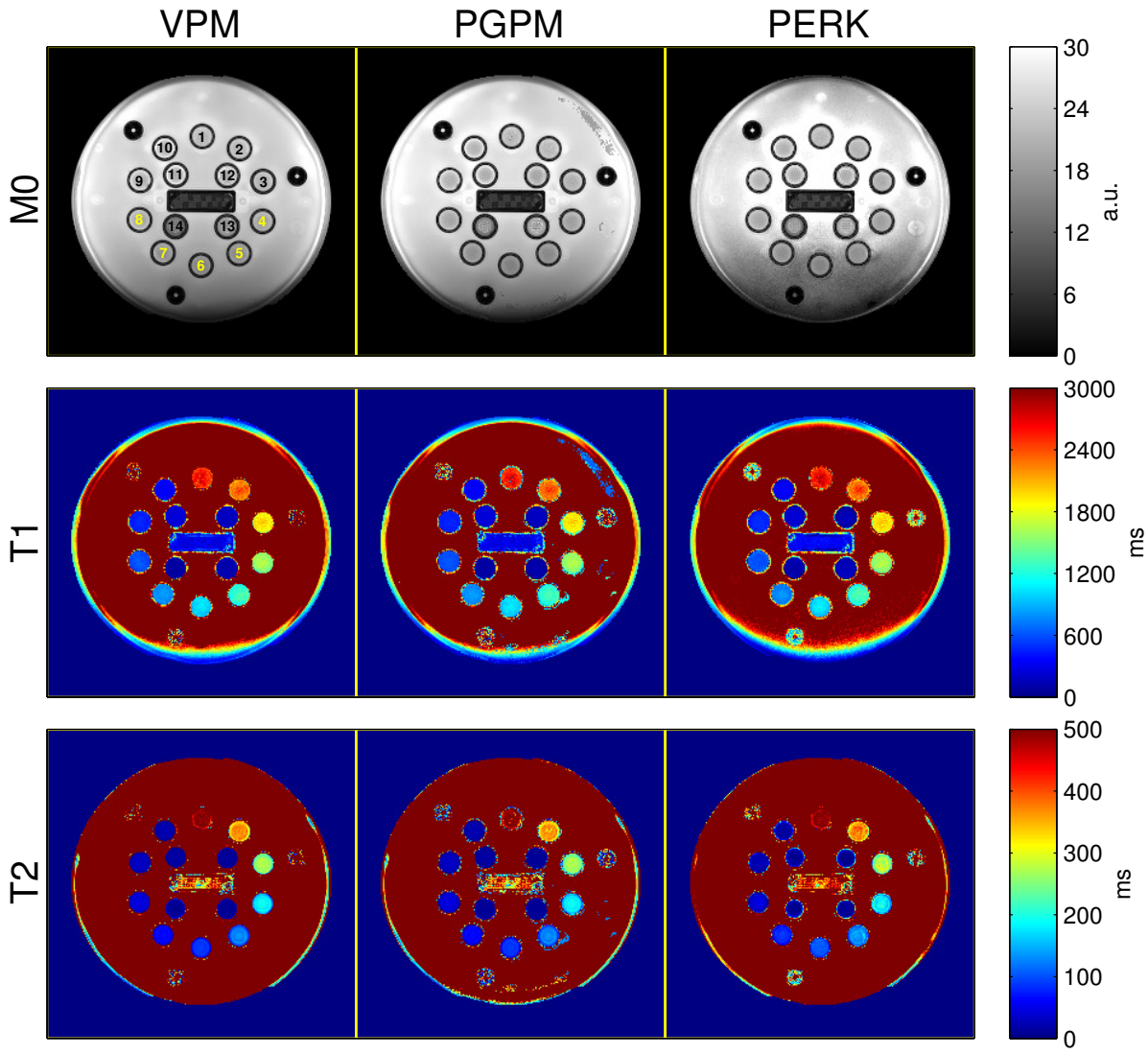


Figure S.9: More aggressively trained VPM, PGPM, and PERK M_0, T_1, T_2 estimates in a quantitative phantom. Here the PERK estimator was trained with a sampling distribution whose support extended over less well identified T_1, T_2 values. Comparing with analogous images in Fig. S.7, PERK performance within vials 4-8 degrades, though in other vials performance clearly improves. Vials are enumerated and highlighted to correspond with markers and colored boxes in Fig. S.8.